Mathematics behind a Class of Image Restoration Algorithms

Luminiţa STATE[1], Cătălina-Lucia COCIANU[2]
[1]University of Piteşti, Piteşti, Romania
[2]Academy of Economic Studies, Bucharest, Romania
radus@sunu.rnc.ro, cocianu@ase.ro

The restoration techniques are usually oriented toward modeling the type of degradation in order to infer the inverse process for recovering the given image. This approach usually involves the option for a criterion to numerically evaluate the quality of the resulted image and consequently the restoration process can be expressed in terms of an optimization problem. Most of the approaches are essentially based on additional hypothesis concerning the statistical properties of images. However, in real life applications, there is no enough information to support a certain particular image model, and consequently model-free developments have to be used instead. In our approaches the problem of image denoising/restoration is viewed as an information transmission/processing system, where the signal representing a certain clean image is transmitted through a noisy channel and only a noise-corrupted version is available. The aim is to recover the available signal as much as possible by using different noise removal techniques that is to build an accurate approximation of the initial image. Unfortunately, a series of image qualities, as for instance clarity, brightness, contrast, are affected by the noise removal techniques and consequently there is a need to partially restore them on the basis of information extracted exclusively from data. Following a brief description of the image restoration framework provided in the introductory part, a PCA-based methodology is presented in the second section of the paper. The basics of a new informational-based development for image restoration purposes and scatter matrix-based methods are given in the next two sections. The final section contains concluding remarks and suggestions for further work.

Keywords: *Principal Component Analysis, Scatter Matrix, Bhattacharyya Upper Margin, Optimal Linear Compression/Decompression, Image Restoration*

1 Introduction

Image restoration methods are used to improve the appearance of an image by the application of a restoration process based on a mathematical model to explain the way the image was distorted by noise. Examples of types of degradation include blurring caused by motion or atmospheric disturbance, geometric distortion caused by imperfect lenses, superimposed interference patterns caused by mechanical systems, and noise induced by electronic sources.

Usually, it is assumed that the degradation model is either known or can be estimated from data. The general idea is to model the degradation process and then apply the inverse process to restore the original image. In cases when the available knowledge does not allow to adopt a reasonable model for the degradation mechanism it becomes necessary to extract information about the noise directed by data and then to use this information for restoration purposes. The knowledge about the particular generation process of the image is application specific. For example, it proves helpful to know how a specific lens distorts an image or how mechanical vibration from a satellite affects an image. This information can be gathered from the analysis of the image acquisition process and by applying image analysis techniques to samples of degraded images.

The restoration can be viewed as a process that attempts to reconstruct or recover a degraded image using some available knowledge about the degradation mechanism. Typically, the noise can be modeled with either a Gaussian, uniform or salt and pepper distribution. The restoration techniques are usually oriented toward modeling the type of

degradation in order to infer the inverse process for recovering the given image. This approach usually involves the option for a criterion to numerically evaluate the quality of the resulted image and consequently the restoration process can be expressed in terms of an optimization problem.

The multi-resolution support set is a data structure suitable for developing noise removal algorithms [2], [3]. The multi-resolution algorithms perform the restoration tasks by combining, at each resolution level, according to a certain rule, the pixels of a binary support image. Some others use a selective wavelet shrinkage algorithm for digital image denoising aiming to improve the performance. For instance Balster [3] proposes an attempt of this sort together with a computation scheme, the denoising methodology incorporated in this algorithm involving a two-threshold validation process for real time selection of wavelet coefficients.

A new solution of the denoising problem based on the description length of the noiseless data in the subspace of the basis is proposed in [4], where the desired description length is estimated for each subspace and the selection of the subspace corresponding to the minimum length is suggested.

In [2], a method for removing Gaussian noise from digital images based on the combination of the wavelet packet transform and the PCA is proposed. The method leads to tailored filters by applying the Karhunen-Loeve transform in the wavelet packet domain and acts with a suitable shrinkage function on these new coefficients, allowing the noise removal without blurring the edges and other important characteristics of the images.

Wavelet thresholding methods modifying the noisy coefficients were proposed by several authors [5], [27]. The attempts are based on the idea that images are represented by large wavelet coefficients that have to be preserved whereas the noise is distributed across the set of small coefficients that have to be canceled. Since the edges lead to a considerable amount of wavelet coefficients of lower values than the threshold, the cancellation of these wavelet coefficients may cause small oscillations near the edges resulting spurious wavelets in the restored image.

Most of the approaches are essentially based on additional hypothesis concerning the statistical properties of images. However, in real life applications, there is no enough information to support a certain particular image model, and consequently model-free developments have to be used instead.

In our approaches the problem of image denoising/restoration is viewed as an information transmission/processing system, where the signal X representing a certain clean image is transmitted through a noisy channel and only a noise-corrupted version $X^{(\eta)}$ is available. The aim is to recover $X^{(\eta)}$ as much as possible by using different noise removal techniques that is to build an accurate approximation of X. Unfortunately, a series of image qualities, as for instance clarity, brightness, contrast, are affected by the noise removal techniques and consequently there is a need to partially restore them on the basis of information extracted exclusively from data. Assume that a noise-removing binomial filter is applied to the output $X^{(\eta)}$ resulting $F\left(X^{(\eta)}\right)$, and $F\left(X^{(\eta)}\right)$ is submitted to a restoration process yielding to $\overline{X}$, an approximation of the X, where the restoration process is based exclusively on $X^{(\eta)}$ and $F\left(X^{(\eta)}\right)$. We assume that, for each image X, a series of its noisy variants $X_1^{(2)},...,X_N^{(2)}$ are available whose corresponding filtered versions are $X_1^{(1)},...,X_N^{(1)}$. As an working assumption, we consider that the noisy variants were generated independently, while the noise that affected the initial clean image X preserved its statistical properties. We consider the normal model for noise, that is $\left\{X_1^{(2)},...,X_N^{(2)}\right\}$ is a Bernoullian sample of the random $r \times c$ -dimensional vector $X^{(\eta)} = X + \eta$, where $\eta \sim N(\mu, \Sigma)$ and $\left\{X_1^{(1)},...,X_N^{(1)}\right\}$ is a sample of the filtered random vector $F\left(X^{(\eta)}\right)$, where $X^{(\eta)}$ and

$F\left(X^{(\eta)}\right)$ are normally distributed. Let us denote by $\mu^{(1)} = E\left(F\left(X^{(\eta)}\right)\right)$, $\mu^{(2)} = E\left(X^{(\eta)}\right)$, Σ_{11}, Σ_{22}, the expectations and the covariance matrices of $F\left(X^{(\eta)}\right)$ and $X^{(\eta)}$ respectively.

In case the assumption that $2\times r\times c$ -dimensional vector $\left(X^{(\eta)}, F\left(X^{(\eta)}\right)\right)$ is normally distributed holds, the conditional distribution of $F\left(X^{(\eta)}\right)$ on $X^{(\eta)}$ is $N\left(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(X^{(\eta)} - \mu^{(2)}\right), \Sigma_{11.2}\right)$, where $E\left(F\left(X^{(\eta)}\right)\middle| X^{(\eta)}\right) = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(X^{(\eta)} - \mu^{(2)}\right)$ is the regression function of $F\left(X^{(\eta)}\right)$ on $X^{(\eta)}$, and $\Sigma_{12} = \mathrm{cov}\left(F\left(X^{(\eta)}\right), X^{(\eta)}\right)$, $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\left(\Sigma_{12}\right)^{T}$.

It is well known that $E\left(F\left(X^{(\eta)}\right)\middle| X^{(\eta)}\right)$ minimizes the variance of $\left(F\left(X^{(\eta)}\right) - Y\right)$ in the class of random vectors $X^{(\eta)}$-measurable and maximizes the correlation coefficient $\rho\left(F\left(X^{(\eta)}\right), Y\right)$ between $F\left(X^{(\eta)}\right)$ and Y in the class of linear functions of $X^{(\eta)}$. Moreover, $E\left(F\left(X^{(\eta)}\right)\middle| X^{(\eta)}\right)$ is $X^{(\eta)}$ -measurable and, since $F\left(X^{(\eta)}\right)$ - $E\left(F\left(X^{(\eta)}\right)\middle| X^{(\eta)}\right)$ and $X^{(\eta)}$ are independent, the whole information carried by $X^{(\eta)}$ with respect to $F\left(X^{(\eta)}\right)$ is contained by $E\left(F\left(X^{(\eta)}\right)\middle| X^{(\eta)}\right)$.

2 Principal Component Analysis (PCA) – based Approaches for Linear Compression/ Decompression

The aim of this section is to present the PCA-based framework for a series of image processing tasks, as noise removal and image recognition. For simplicity sake, we assume that the image is represented by a n-dimensional real-valued random vector *X* of mean 0 and covariance matrix Σ. The main idea is that, as it is often experimentally confirmed, the minor components of the repartition of *X* contain the maximum amount of noise. Consequently, a possible strategy to clean the image is to filter it by applying a compression process that yields to a representation of the image in the space of major components followed by a lifting or decompression step that produces cleaned versions of *X* in the initial space. Our approach in designing the compression/decompression filters is of linear type, that is both modules are modeled in terms of linear transforms. If we denote by *m* the desired dimension of the feature space, $1 \le m \le n$, and let A and B be the linear compression, and decompression filters respectively, that is the entries of the m-dimensional representation of *Y* are linear combinations of the entries of X, and the columns of A can be viewed as feature vectors (Figure 1). Since the compression should minimized the loss of information, usually the design of the compression filters aims to remove redundancy from the processed image *X*. Consequently, the feature vectors should be at least linearly independent, that is a natural requirement is to assume $rank(A) = m$. The most popular criterion function to express the quality of a linear compression/decompression scheme is the mean error $\varepsilon^2(m, A, B) = E\left(\left\|X - \hat{X}\right\|^2\right)$ and, for given *m*, the optimal linear compression/decompression scheme $\left(A^*, B^*\right)$ minimizes $\varepsilon^2(m, A, B)$, that is $\left(A^*, B^*\right) = \arg \min_{\substack{A\in M_{nxm}\\ rank(A)=m\\ B\in M_{nxm}}} \varepsilon^2(m, A, B)$. This optimization problem can be solved by decomposing it into two simpler optimization problems, namely,

P1: For given $A \in M_{nxm}$ such that $rank(A) = m$, compute $B^*(A) = \arg \min_{B\in M_{nxm}} \varepsilon^2(m, A, B)$.

P2: Compute $A^* = \arg \min_{\substack{A\in M_{nxm}\\ rank(A)=m}} \varepsilon^2\left(m, A, B^*(A)\right)$

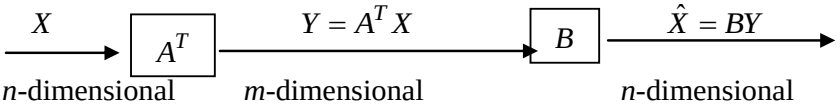

Fig. 1. The solutions of P1 and P2 are given by the following theorems [30]

Theorem 1. The solution of P1 is $B^*(A) = SA\left(A^T SA\right)^{-1}$, where S is the autocorrelation matrix of the repartition of X.

Theorem 2. Let be the $\theta_1 \geq \theta_2 \geq ... \geq \theta_m \geq ... \geq \theta_n > 0$ eigenvalues of S and $\Phi_1, \Phi_2, ..., \Phi_n$ unit orthogonal corresponding eigenvectors; $\Phi^{(m)} = [\Phi_1, \Phi_2, ..., \Phi_m]$. The set of solutions of P2 is $\left\{A^* = \Phi^{(m)} * C \middle| C \text{ orthogonal}\right\}$.

Moreover, $\varepsilon^2\left(m, A^*, B^*\right) = \sum_{i=m+1}^{n} \theta_i$.

Note that in case $C = I_m$, an optimal linear compression/decompression scheme is $\left(A^*, B^*\right) = \left(\Phi^{(m)}, \Phi^{(m)}\right)$.

The principal directions of the repartition of X are the directions corresponding to the maximum variability, where the variability is expressed in terms of the variance.

Definition. The vector $\Psi_1 \in \mathbf{R}^n$ is the first principal direction if $\|\Psi_1\| = 1$ and $\text{var}\left(\Psi_1^T X\right) = \sup_{\substack{\Phi \in R^n \\ \|\Phi\|=1}} \text{var}\left(\Phi^T X\right)$.

The value $\Psi_1^T X$ is referred as the first principal component of X.

Now, recursively, for any k, $2 \leq k \leq n$, if we denote by $L^{\perp}(\Psi_1, ..., \Psi_{k-1})$ the linear subspace orthogonal on the linear subspace generated by the first (k-1) directions, $\Psi_k \in \mathbf{R}^n$ is a k-th principal direction if $\|\Psi_k\| = 1$ and $\text{var}\left(\Psi_k^T X\right) = \sup_{\substack{\Phi \in L^{\perp}(\Psi_1, ..., \Psi_{k-1}) \\ \|\Phi\|=1}} \text{var}\left(\Phi^T X\right)$.

The value $\Psi_k^T X$ is referred as the k-th principal component of the signal X.

Note that a set of principal directions $\Psi_1, ..., \Psi_n$ of the repartition of X is an orthogonal basis of $\mathbf{R}^n$, and $Y = \Psi^T X$ being the image representation in terms of the selected principal directions, where $\Psi = [\Psi_1, ..., \Psi_n]$.

The fundamental result is given by the celebrated Karhunen-Loeve theorem.

Theorem 3. Let X be a n-dimensional real-valued random vector and $Cov\left(X, X^T\right) = \Sigma$. Then the principal directions $\Psi_1, ..., \Psi_n$ are orthogonal unit eigenvectors of Σ.

If we denote by $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ the eigenvalues of Σ, then, for any k, $1 \leq k \leq n$, the k-th principal direction is an eigenvector of Σ associated to λ_k.

In case we use the orthogonal basis $\Psi_1, ..., \Psi_n$ the representation of X is $X = \sum_{i=1}^{n} y_i \psi_i$, where $y_i = \psi_i{}^T X$, $1 \leq i \leq n$. Obviously, $\text{var}(y_i) = \psi_i^T E\left((X - \mu)(X - \mu)^T\right)\psi_i = \psi_i^T \Sigma \psi_i = \lambda_i$, where μ is the mean vector of the repartition of X. Assume that the linear compression/decompression scheme is based on the information contained by the linear subspace spanned by the principal directions, the compression/decompression filters are both, $\psi^{(m)} = [\Psi_1, ..., \Psi_m]$. Then the mean error becomes $\varepsilon^2\left(m, \psi^{(m)}, \psi^{(m)}\right) = \sum_{i=m+1}^{n} \lambda_i$.

Note that in case of 0-mean repartitions, the autocorrelation matrix S equals the covariance matrix Σ, that is $S = \Sigma$. In this case, according to the results given by Theorem 2 and Theorem 3, the optimal linear compression/ decompression scheme from the point of view of mean error criterion is based

exclusively on the information contained by the linear *m*-dimensional subspace of principal directions that is the subspace where the repartition of *X* is of maximum variability.

The sum $\sum_{i=m+1}^{n}\lambda_i$ is a measure of the amount of information lost because of ignoring the minor components $\Psi_{m+1},...,\Psi_n$ in the sense of the mean error criterion.

Although, in the general case when the $\mu \neq O_n$, the principal directions are different from the columns of the optimal linear compression filter, experimentally it is confirmed that both subspaces $L^{\perp}(\Phi_1,\Phi_2,...,\Phi_m)$ and $L^{\perp}(\Psi_1,...,\Psi_m)$ contain the maximum amount of noise, that is put in other words the compressed/decompressed images $\tilde{X} = \Phi^{(m)}\left(\Phi^{(m)}\right)^T X$ and $\hat{X} = \psi^{(m)}\left(\psi^{(m)}\right)^T X$ correspond to cleaned versions of *X*.

In practical applications, the statistical properties of the processed images are unknown, but a series of versions $X_1, X_2,..., X_N$ are available instead, we use as an working assumption the hypothesis that they corresponds to an i.i.d. sample from the repartition of *X* and the principal directions are estimated from data. The accuracy of the resulted linear compression/decompression scheme depends essentially on the size *N* of the sample, but this dependency is extremely hard or even impossible to be evaluated in advance. The alternative is to gradually improve the accuracy by taking more examples, that is to enlarge the sample size, but the re-computation of the estimates of the covariance matrices, autocorrelation matrices, eigenvalues and eigenvectors is computationally complex. Hopefully, the re-computation can be avoided using the first order approximation schemes.

The estimates of the covariance matrices and autocorrelation matrices on the basis of the samples $X_1, X_2,..., X_N$, are

$\hat{\Sigma}_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu}_N)(X_i - \hat{\mu}_N)^T$, and

$\hat{S}_N = \hat{\Sigma}_N + \hat{\mu}_N(\hat{\mu}_N)^T$, respectively,

where $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$.

The estimates computed on the basis of $\{X_1, X_2,..., X_N\} \cup \{X_{N+1}\}$ can be simply re-computed using $\hat{\mu}_N$, $\hat{\Sigma}_N$, $\hat{S}_N$, and X_{N+1} as follows.

$$\hat{\mu}_{N+1} = \frac{N}{N+1}\hat{\mu}_N + \frac{1}{N+1}X_{N+1}$$

$$\hat{\Sigma}_{N+1} = \frac{N-1}{N}\hat{\Sigma}_N + \frac{1}{N+1}(X_{N+1} - \hat{\mu}_N)(X_{N+1} - \hat{\mu}_N)^T$$

$$\hat{S}_{N+1} = \hat{\Sigma}_{N+1} + \hat{\mu}_{N+1}(\hat{\mu}_{N+1})^T$$

Unfortunately, the expressions of the eigenvalues and eigenvectors of $\hat{\Sigma}_{N+1}$, and $\hat{S}_{N+1}$ in terms of the eigenvalues and eigenvectors of $\hat{\Sigma}_N$, and $\hat{S}_N$ respectively cannot be directly derived and consequently we are forced to derive first-order approximations instead.

The following lemma establishes first order approximations for the eigenvalues and eigenvectors of the matrix $\hat{\Sigma}_{N+1}$, and approximations of the eigenvalues and eigenvectors of $\hat{S}_{N+1}$ can be obtained using similar computations.

Let us denote by $\lambda_1^N \geq \lambda_2^N \geq ... \geq \lambda_n^N$ the eigen values and by $\psi_1^N,...,\psi_n^N$ a set of orthonormal eigen vectors of $\hat{\Sigma}_N$.

Lemma 1. In case the eigen values of $\hat{\Sigma}_N$ are pairwise distinct, the following first order approximations hold [33],

$$\lambda_i^{N+1} = \lambda_i^N + \left(\psi_i^N\right)^T \Delta\hat{\Sigma}_N \psi_i^N = \left(\psi_i^N\right)^T \hat{\Sigma}_{N+1}\psi_i^N$$

$$\psi_i^{N+1} = \psi_i^N + \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left(\psi_N^j\right)^T \Delta\hat{\Sigma}_N \psi_i^N}{\lambda_i^N - \lambda_j^N}\psi_j^N$$

Proof. Using the perturbation theory, we get,

$\hat{\Sigma}_{N+1} = \hat{\Sigma}_N + \Delta\hat{\Sigma}_N$ and,
$\psi_i^{N+1} = \psi_i^N + \Delta\psi_i^N$, $\lambda_i^{N+1} = \lambda_i^N + \Delta\lambda_i^N$, $1 \le i \le n$.
Then,

$$\Delta\hat{\Sigma}_N = \frac{1}{N+1}(X_{N+1} - \mu_N)(X_{N+1} - \mu_N)^T - \frac{1}{N}\hat{\Sigma}_N$$

$$\left(\hat{\Sigma}_N + \Delta\hat{\Sigma}_N\right)\left(\psi_i^N + \Delta\psi_i^N\right) = \left(\lambda_i^N + \Delta\lambda_i^N\right)\left(\psi_i^N + \Delta\psi_i^N\right)$$

Using first-order approximations, we get,

$$\lambda_i^N\psi_i^N + \hat{\Sigma}_N\Delta\psi_i^N + \Delta\hat{\Sigma}_N\psi_i^N \cong \lambda_i^N\psi_i^N + \lambda_i^N\Delta\psi_i^N + \Delta\lambda_i^N\psi_i^N$$

hence,

$$\left(\psi_i^N\right)^T\hat{\Sigma}_N\Delta\psi_i^N + \left(\psi_i^N\right)^T\Delta\hat{\Sigma}_N\psi_i^N \cong \lambda_i^N\left(\psi_i^N\right)^T\Delta\psi_i^N + \Delta\lambda_i^N\left\|\psi_i^N\right\|^2$$

Using $\lambda_i^N\left(\psi_i^N\right)^T = \left(\psi_i^N\right)^T\hat{\Sigma}_N$ we obtain ,

$$\lambda_i^N\left(\psi_i^N\right)^T\Delta\psi_i^N + \left(\psi_i^N\right)^T\Delta\hat{\Sigma}_N\psi_i^N \cong \lambda_i^N\left(\psi_i^N\right)^T\Delta\psi_i^N + \Delta\lambda_i^N$$

hence $\Delta\lambda_i^N = \left(\psi_i^N\right)^T\Delta\hat{\Sigma}_N\psi_i^N$ that is,

$$\lambda_i^{N+1} = \lambda_i^N + \left(\psi_i^N\right)^T\Delta\hat{\Sigma}_N\psi_i^N = \left(\psi_i^N\right)^T\hat{\Sigma}_{N+1}\psi_i^N$$

The first order approximations of the orthonormal eigenvectors of $\hat{\Sigma}_{N+1}$ can be derived using the expansion of each vector $\Delta\psi_i^N$ in the basis represented by the orthonormal eigen vectors of $\hat{\Sigma}_N$,

$$\Delta\psi_i^N = \sum_{j=1}^{n} b_{i,j}\psi_j^N ,$$

where $b_{i,j} = \left(\psi_j^N\right)^T\Delta\psi_i^N$.
Using the orthonormality, we get,

$$1 = \left\|\psi_i^N + \Delta\psi_i^N\right\|^2 \cong \left\|\psi_i^N\right\|^2 + 2\left(\psi_i^N\right)^T\left(\Delta\psi_i^N\right) = 1 + 2\left(\psi_i^N\right)^T\left(\Delta\psi_i^N\right)$$

that is $b_{i,i} = \left(\psi_i^N\right)^T\Delta\psi_i^N = 0$
The approximation,
$\hat{\Sigma}_N\Delta\psi_i^N + \Delta\hat{\Sigma}_N\psi_i^N \cong \lambda_i^N\Delta\psi_i^N + \Delta\lambda_i^N\psi_i^N$.
holds for each $1 \le i \le n$.
For $1 \le j \ne i \le n$, we obtain the following equations,

$$\left(\psi_N^j\right)^T\hat{\Sigma}_N\Delta\psi_i^N + \left(\psi_N^j\right)^T\Delta\hat{\Sigma}_N\psi_i^N \cong \lambda_i^N\left(\psi_N^j\right)^T\Delta\psi_i^N + \Delta\lambda_i^N\left(\psi_N^j\right)^T\psi_i^N$$

$$\left(\psi_N^j\right)^T\hat{\Sigma}_N\Delta\psi_i^N + \left(\psi_N^j\right)^T\Delta\hat{\Sigma}_N\psi_i^N \cong \lambda_i^N\left(\psi_N^j\right)^T\Delta\psi_i^N$$

We get,

$$\left(\lambda_i^N - \lambda_j^N\right)\left(\psi_N^j\right)^T\Delta\psi_i^N = \left(\psi_N^j\right)^T\Delta\hat{\Sigma}_N\psi_i^N$$

$$b_{i,j} = \left(\psi_j^N\right)^T\Delta\psi_i^N = \frac{\left(\psi_N^j\right)^T\Delta\hat{\Sigma}_N\psi_i^N}{\lambda_i^N - \lambda_j^N}$$

Consequently, the first-order approximation of the eigenvectors of $\hat{\Sigma}_{N+1}$ are,

$$\psi_i^N + \Delta\psi_i^N = \psi_i^N + \sum_{\substack{j=1 \\ j\ne i}}^{n}\frac{\left(\psi_N^j\right)^T\Delta\hat{\Sigma}_N\psi_i^N}{\lambda_i^N - \lambda_j^N}\psi_j^N$$

On the other hand, when an object has to be removed from the sample, then the estimate of the covariance matrix can be computed as,
$\hat{\Sigma}_N = \hat{\Sigma}_{N+1} + \Delta\hat{\Sigma}_{N+1}$,
where

$$\Delta\hat{\Sigma}_{N+1} = \frac{1}{N-1}\hat{\Sigma}_{N+1} - \frac{N}{(N-1)(N+1)}(X_{N+1} - \mu_N)(X_{N+1} - \mu_N)^T$$

and

$$\mu_N = \frac{(N+1)\mu_{N+1}}{N} - \frac{X_{N+1}}{N}$$

3 Information-based approaches for image restoration purposes

The basics of the informational-based method for image restoration purposes are given by the following theoretical results [29].
If X is a continuous n-dimensional random vector and f is the density function of X repartition, we denote by $H(X) = -\int_{R^n} f(x)\ln f(x)dx$ the differential entropy (Shannon) of X.

Lemma 2 Let X be a continuous n-dimensional random vector and $A \in M_n(R)$ a non-singular matrix, $Y = AX$. Then,
$H(X) = H(Y) - \ln|A|$.

Proof. If we denote by g the density function of the random vector Y , then

$g(y) = f\left(A^{-1}y\right)\frac{1}{|A|}$.

Using straightforward computation, we get,

$H(X) = -\int_{R^n} f(x)\ln f(x)dx =$

$-\frac{1}{|A|}\int_{R^n} f\left(A^{-1}y\right)\ln f\left(A^{-1}y\right)dy =$

$-\frac{1}{|A|}\int_{R^n}|A|g(y)\ln\left(|A|g(y)\right)dy =$

$-\int_{R^n} g(y)\ln(g(y))dy - \ln|A|\left(\int_{R^n} g(y)dy\right) = H(Y) - \ln|A|$

Lemma 3. Let X be a continuous n-dimensional normally distributed random vector, $X \sim N(0,\Sigma)$ and q a given natural number, $1 \le q < n$. If $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ where $X^{(1)}$ is q-dimensional, then, for any $x^{(2)} \in \mathbf{R}^{n-q}$,

$H\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right) = H\left(X^{(1)} - E\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)\right)$,

where $E\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)$ is the regression function of $X^{(1)}$ on $X^{(2)} = x^{(2)}$, and $H\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)$ is the conditional differential entropy of $X^{(1)}$ being given $X^{(2)} = x^{(2)}$.

Proof. Let $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$

where $\Sigma_{11} = \mathrm{cov}\left(X^{(1)}, X^{(1)T}\right)$,

$\Sigma_{22} = \mathrm{cov}\left(X^{(2)}, X^{(2)T}\right)$,

$\Sigma_{12} = \mathrm{cov}\left(X^{(1)}, X^{(2)T}\right)$.

Since $E(X) = 0$, we get

$E\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right) = \Sigma_{12}\Sigma_{22}^{-1}x^{(2)}$.

The linear transform of matrix $A = \begin{pmatrix} I_q & -\Sigma_{12}\Sigma_{22} \\ 0 & I_{n-q} \end{pmatrix}$ de-correlates $X^{(1)}, X^{(2)}$ and consequently $Y^{(1)}$, $Y^{(2)}$ are normally distributed independent random vectors [1], therefore $H(Y) = H\left(Y^{(1)}\right) + H\left(Y^{(2)}\right)$,

where $Y = AX = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix}$.

Moreover,

$Y^{(1)} = X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)} = X^{(1)} - E\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)$, $Y^{(2)} = X^{(2)}$, and $Y^{(1)} \sim N\left(0,\Sigma_{11.2}\right)$, $Y^{(2)} \sim N\left(0,\Sigma_{22}\right)$,

where $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\left(\Sigma_{12}\right)^T$.

Since $|A| = 1$, using Lemma 2, we get

$H(X) = H(Y)$, that is

$H\left(X^{(1)}, X^{(2)}\right) = H\left(X^{(1)} - E\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)\right) + H\left(X^{(2)}\right)$

Finally, taking into account that $H\left(X^{(1)}, X^{(2)}\right) = H\left(X^{(2)}\right) + H\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)$, we obtain

$H\left(X^{(1)} - E\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)\right) = H\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)$

Since $H\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)$ represents a measure of the amount of incertitude still remaining with respect to $X^{(1)}$ when $X^{(2)}$ is known, the previous relation establishes that the whole information contained by $X^{(2)}$ with respect to $X^{(1)}$ is concentrated on $E\left(X^{(1)}\middle|X^{(2)} = x^{(2)}\right)$.

Note that since the entropy of the n-dimensional normal distribution $N(\mu,\Sigma)$ is given by $H(X) = \frac{n}{2}\ln 2\pi e + \frac{1}{2}\ln|\Sigma|$, we get

$H\left(X^{(1)} - E\left(X^{(1)}\middle|X^{(2)}\right)\right) = H\left(X^{(1)}\middle|X^{(2)}\right) = \frac{q}{2}\ln 2\pi e + \frac{1}{2}\ln|\Sigma_{11.2}|$.

Using the notations introduced in Section 1 and the results established in lemmas 2 and 3, we obtain,

$H\left(F\left(X^{(\eta)}\right) - E\left(F\left(X^{(\eta)}\right)\middle|X^{(\eta)}\right)\right) = H\left(F\left(X^{(\eta)}\right)\middle|X^{(\eta)}\right)$

and conclude that $E\left(F\left(X^{(\eta)}\right)\middle|X^{(\eta)}\right)$ contains the whole information existing in $X^{(\eta)}$ with respect to $F\left(X^{(\eta)}\right)$ a part of it being responsible of the initial existing noise η and another component being responsible of the degradation of quality .

4 The image restoration method based on scatter matrices and on bounds on the probability of error

In statistical discriminant analysis, within-class, between-class and mixture scatter matrices are used to formulate criteria of class separability.

In case we need to discriminate between *m* classes $H_i, i=1,m$ and $\left\{X_1^{(i)},...,X_N^{(i)}\right\}$ are examples of patterns coming respectively from these classes, the within -class scatter matrix shows the scatter of samples around their class expected vectors and it is typically given by the expression $S_w=\sum_{i=1}^{m}\xi_i\sum_{k=1}^{N}\left(X_k^{(i)}-\hat{\mu}_i\right)\left(X_k^{(i)}-\hat{\mu}_i\right)^T$, where $\hat{\mu}_i$ is the prototype of H_i and ξ_i is the *a priori* probability of $H_i, i=1,m$.

Very often, the *a priori* probabilities are taken $\xi_i=\frac{1}{m}$ and each prototype is computed as the weighted mean of the patterns belonging to the respective class.

The between-class scatter matrix is the scatter of the expected vectors around the mixture mean as $S_b=\sum_{i=1}^{m}\xi_i\sum_{k=1}^{N}\left(\hat{\mu}_i-\mu_0\right)\left(\hat{\mu}_i-\mu_0\right)^T$ where μ_0 represents the expected vector of the mixture distribution; usually μ_0 is taken as $\mu_0=\sum_{i=1}^{m}\xi_i\hat{\mu}_i$.

The mixture scatter matrix is the covariance matrix of all samples regardless of their class assignments and it is defined by $S_m=S_w+S_b$. Note that all these scatter matrices are designed to be invariant under coordinate shifts.

In order to formulate criteria for class separability, these matrices should be converted into a number. This number should be larger when the between-class scatter is larger or the within-class scatter is smaller. Typical criteria are $J_1=tr\left(S_2^{-1}S_1\right)$, $J_2=\ln\left|S_2^{-1}S_1\right|$, where $(S_1,S_2)\in\{(S_b,S_w),(S_b,S_m),(S_w,S_m),(S_m,S_w)\}$ and their values can be taken as measures of overall class separability. Obviously, both criteria are invariant under linear non-singular transforms and they are currently used for feature extraction purposes. When the linear feature extraction problem is solved on the basis of either J_1 or J_2, their values are taken as numerical indicators of the loss of information implied by the reduction of dimensionality and implicitly deteriorating class separability. Consequently, the best linear feature extraction is formulated as the optimization problem $\arg\left(\inf_{A\in R^{n*m}}\left|J_k(m,A)-J_k\right|\right)$ where m stands for the desired number of features , $J_k(m,A)$is the value of the criterion $J_k, k=1,2$ in the transformed m-dimensional space of $Y=A^TX$, where A is a $n*m$ matrix .

If the pattern classes are represented by the noisy image $X^{(\eta)}$ and the filtered image $F\left(X^{(\eta)}\right)$ respectively, the value of each of the criteria $J_k, k=1,2$ is a measure of overall class separability as well as a measure of the information discriminating between these classes. In other words, $J_k, k=1,2$ can be taken as measuring the effects of the noise removing filter expressing a measure of the quantity of information lost due to the use of the particular filter. In order to remove at least partially

The idea of our attempt is to use the most informative features discriminating between $X^{(\eta)}$ and $F\left(X^{(\eta)}\right)$ for getting correction terms in restoring the filtered images $F\left(X^{(\eta)}\right)$. The attempt is justified by the argument that besides information about the removed noise, the most informative features discriminating

between $X^{(\eta)}$ and $F\left(X^{(\eta)}\right)$ would contain appropriate information allowing the quality improvement of the image $F\left(X^{(\eta)}\right)$. Let $\left\{X_1^{(2)},...,X_N^{(2)}\right\}$ be the sample of noise corrupted versions of the $r\times c$-dimensional image X and $\left\{X_1^{(1)},...,X_N^{(1)}\right\}$ their filtered versions, $X_i^{(1)}=F\left(X_i^{(2)}\right)$, $i=1,N$. We assume $\xi_1=\xi_2=0.5$, therefore the scatter matrices become $S_w=\hat{\Sigma}_1+\hat{\Sigma}_2$,

$S_b=\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)^T$ and

$S_m=S_w+S_b$ where $\hat{\mu}^{(i)}=\frac{1}{N}\sum_{k=1}^{N}X_k^{(i)}$,

$\hat{\Sigma}_i=\frac{1}{N-1}\sum_{k=1}^{N}\left(X_k^{(i)}-\hat{\mu}^{(i)}\right)\left(X_k^{(i)}-\hat{\mu}^{(i)}\right)^T$, $i=1,2$

Since $rank\left(S_b\right)=1$, we get $rank\left(S_w^{-1}S_b\right)=1$, that is when $S_2=S_w$ and $S_1=S_b$,the matrix $S_2^{-1}S_1$ has an unique positive eigenvalue, its unit eigenvector being given by

$$\Phi_1=\frac{S_w^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)}{\left\|S_w^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)\right\|}$$

Lemma 4. For any m, $1\le m\le n$,

$$\arg\left(\inf_{A\in R^{n*m}}\left|J_k(m,A)-J_k\right|\right)=\left\{A\Psi\middle|A=(\Phi_1,\Phi_2,...,\Phi_m),\Psi\in R^{m*m},\left|\Psi\right|\neq 0\right\}$$

where $\Phi_1,...,\Phi_m$ are unit eigenvectors corresponding to the m largest eigenvalues of $S_2^{-1}S_1$ [10].

Proof. Let C be the matrix diagonalizing simultaneously the symmetric matrices S_1, S_2, $C^TS_2C=I_n$, $C^TS_1C=\Lambda$, $CC^T=S_2^{-1}$, where $\Lambda=diag\left(\lambda_1,...,\lambda_n\right)$.

It is well known that the eigenvalues $\lambda_1,\lambda_2,...,\lambda_n$, of $S_2^{-1}S_1$ are real positive values, and the columns of C are eigenvectors of $S_2^{-1}S_1$. Therefore,

$$J_1=tr\left(S_2^{-1}S_1\right)=tr\left(CC^TS_1\right)=tr\left(C^TS_1C\right)=\sum_{k=1}^{n}\lambda_k$$

$$J_2=\ln\left|S_2^{-1}S_1\right|=\ln\left|CC^TS_1\right|=\ln\left|\Lambda\right|=\sum_{k=1}^{n}\ln\lambda_k$$

Let $A\in M_{n\times m}(R)$ be such that $rank(A)=m$ and $Y=A^TX$; we denote by $S_1(m,A)$, $S_2(m,A)$ the counterparts of S_1, S_2 in the m-dimensional resulted space. Obviously, for

$(S_1,S_2)\in\left\{(S_b,S_w),(S_b,S_m),(S_w,S_m),(S_m,S_w)\right\}$

we get $S_k(m,A)=A^TS_kA$, $k=1,2$

and consequently,

$J_1(m,A)=tr\left(\left(A^TS_2A\right)^{-1}A^TS_1A\right)=tr\left(\Lambda(m,A)\right)$,

where $\Lambda(m,A)$ is the diagonal matrix having as entries the eigenvalues of $S_2^{-1}(m,A)S_1(m,A)$.

The critical points of $J_1(m,A)$ are the matrices A that are solutions of the equation $\frac{\partial}{\partial A}J_1(m,A)=0$, that is

$S_2AS_2^{-1}(m,A)S_1(m,A)S_2^{-1}(m,A)=S_1AS_2^{-1}(m,A)$

which can be also written as $S_2^{-1}S_1A=AS_2^{-1}(m,A)S_1(m,A)$.

Let $B\in M_m(R)$ be the matrix diagonalizing simultaneously $S_1(m,A)$, $S_2(m,A)$.

Since $B^TS_2(m,A)B=I_m$,

$B^TS_1(m,A)B=\Lambda(m,A)$

and $BB^T=S_2^{-1}(m,A)$, we obtain that A is a critical point of $J_1(m,A)$ if and only if $S_2^{-1}S_1AB=AB\Lambda(m,A)$ that is, if and only if the columns of AB are eigenvectors and the entries of the diagonal matrix $\Lambda(m,A)$ are the corresponding eigenvalues of $S_2^{-1}S_1$. Consequently, $J_1(m,A)\le J_1$ and $\left|J_1(m,A)-J_1\right|$ is minimized when the linear feature extractor A is such that the columns of AB are the eigenvectors corresponding to the largest m eigenvalues of $S_2^{-1}S_1$ .Obviously the criterion function J_1 is invariant with respect to non-singular transforms therefore $\Phi^{(m)}=(\Phi_1,...,\Phi_m)$ can be taken as the optimal linear feature extractor where $\Phi_i, i=1,m$ are the eigenvectors corresponding to the largest eigenvalues of $S_2^{-1}S_1$.

In case the criterion function J_2 is selected to measure the effects of the noise removing filter, the critical points of $J_2(m,A)=\ln\left|S_2^{-1}(m,A)S_1(m,A)\right|=$ $\ln\left|A^TS_1A\right|-\ln\left|A^TS_2A\right|$ are the solutions of the equation $\frac{\partial}{\partial A}J_1(m,A)=0$. By straightforward computation we obtain $S_2AS_2^{-1}(m,A)=S_1AS_1^{-1}(m,A)$, that is $\left(S_2^{-1}S_1\right)A=A\left(S_2^{-1}(m,A)S_1^{-1}(m,A)\right)$.

Therefore, the optimal linear feature extractor from the point of view of both criteria J_k, $k=1,2$ is given by $\Phi^{(m)}=(\Phi_1,...,\Phi_m)$.

The probability of error is the most effective measure of classification decision rule usefulness, but its evaluation involves integrations on complicated regions in high dimensional spaces. When a closed-form expression for the error probability cannot be obtained, we may seek either for approximate expressions, or upper/lower bounds for the error probability.

Assume that the design of the Bayes classifier is intended to discriminate between two pattern classes and the available information is represented by mean vectors μ_i, $i=1,2$ and the covariance matrices Σ_i, $i=1,2$ corresponding to the repartitions of the classes respectively. The Chernoff upper bounds of the Bayesian error [15] are given by $\varepsilon_s=\xi_1^s\xi_2^{1-s}\int\left(f_1(x)\right)^s\left(f_2(x)\right)^{1-s}dx$, $s\in[0,1]$, where $\xi=(\xi_1,\xi_2)$ is the *a priori* distribution and f_i is the density function corresponding to the $i-th$ class, $i=1,2$. When both density functions are normal, $f_i\sim N(\mu_i,\Sigma_i)$ $i=1,2$, the integration can be carried out to obtain a closed-form expression for ε_s, that is $\int\left(f_1(x)\right)^s\left(f_2(x)\right)^{1-s}dx=\exp(-\mu(s))$ where $\mu(s)=\frac{s(1-s)}{2}(\mu_2-\mu_1)^T\left(s\Sigma_1+(1-s)\Sigma_2\right)^{-1}(\mu_2-\mu_1)+\frac{1}{2}\ln\frac{\left|s\Sigma_1+(1-s)\Sigma_2\right|}{\left|\Sigma_1\right|^s\left|\Sigma_2\right|^{1-s}}$

The upper bound $\mu\left(\frac{1}{2}\right)=\frac{1}{8}(\mu_2-\mu_1)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_2-\mu_1)+\frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{\left|\Sigma_1\right|\left|\Sigma_2\right|}}$ is called the Bhattacharyya distance and it is frequently used as a measure of the separability between two repartitions. Using straightforward computations, the Bhattacharyya distance can be written as, $\mu\left(\frac{1}{2}\right)=\frac{1}{8}tr\left\{\overline{\Sigma}^{-1}(\mu_2-\mu_1)(\mu_2-\mu_1)^T\right\}+\frac{1}{4}\ln\left|2I_n+\Sigma_1\Sigma_2^{-1}+\Sigma_2\Sigma_1^{-1}\right|-\frac{n}{4}\ln 2$

where $\overline{\Sigma}=\frac{\Sigma_1+\Sigma_2}{2}$.

Note that one of the first two terms of $\mu\left(\frac{1}{2}\right)$ vanishes, when $\mu_1=\mu_2$, $\Sigma_1=\Sigma_2$ respectively, that is the first term expresses the class separability due to the mean-difference while the second one gives the class separability due to the covariance difference.

The Bhattacharyya distance can be used as criterion function as well to express the quality of a linear feature extractor of matrix $A\in R^{nxm}$. When $\Sigma_1=\Sigma_2=\Sigma$, $J=\mu\left(\frac{1}{2}\right)=\frac{1}{8}tr\left\{\overline{\Sigma}^{-1}(\mu_2-\mu_1)(\mu_2-\mu_1)^T\right\}$ therefore J is a particular case of the criterion J_1 for $S_2=\Sigma$ and $S_1=S_b=(\mu_2-\mu_1)(\mu_2-\mu_1)^T$. Consequently the whole information about the class separability is contained by an unique feature $\Phi_1=\frac{\Sigma^{-1}(\mu_2-\mu_1)}{\left\|\Sigma^{-1}(\mu_2-\mu_1)\right\|}$. When $\mu_1=\mu_2$ and $\Sigma_1\neq\Sigma_2$,

$J = \frac{1}{4}\ln\left|2I_n + \Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2\right| - \frac{n}{4}\ln 2 = \frac{1}{4}\sum_{j=1}^{n}\left(2 + \lambda_j + \frac{1}{\lambda_j}\right) - \frac{n}{4}\ln 2$

where $\lambda_j, j = 1, n$ are the eigenvalues of $\Sigma_1^{-1}\Sigma_2$.

If the linear feature extractor is defined by the matrix $A \in R^{nxm}$, then the value of the Bhattacharyya distance in the transformed space $Y = A^T X$ is given by, $J(m, A) = \frac{1}{4}\ln\left|2I_m + \left(A^T\Sigma_2 A\right)^{-1}\left(A^T\Sigma_1 A\right) + \left(A^T\Sigma_1 A\right)^{-1}\left(A^T\Sigma_2 A\right)\right| - \frac{m}{4}\ln 2$.

The critical points of $J(m, A)$ are the solutions of the equation $\frac{\partial J(m,A)}{\partial A} = 0$ that is,

$B\left\{\Sigma_2 A\Sigma_2^{-1}(m)\Sigma_1(m)\Sigma_2^{-1}(m) - \Sigma_1 A\Sigma_2^{-1}(m)\right\} + B\left\{\Sigma_1 A\Sigma_2^{-1}(m)\Sigma_2(m)\Sigma_1^{-1}(m) - \Sigma_2 A\Sigma_1^{-1}(m)\right\} = 0$

where $\Sigma_i(m) = A^T\Sigma_i A$, $i = 1,2$ and $B = \left[\left(A^T\Sigma_1 A\right)^{-1}\left(A^T\Sigma_2 A\right) + \left(A^T\Sigma_2 A\right)^{-1}\left(A^T\Sigma_1 A\right) + 2I_m\right]^{-1}$

Suboptimal solutions can be identified as the solutions of the system

$$\begin{cases}\Sigma_2 A\Sigma_2^{-1}(m)\Sigma_1(m)\Sigma_2^{-1}(m) - \Sigma_1 A\Sigma_2^{-1}(m) = 0\\ \Sigma_1 A\Sigma_1^{-1}(m)\Sigma_2(m)\Sigma_1^{-1}(m) - \Sigma_2 A\Sigma_1^{-1}(m) = 0\end{cases}$$

or equivalently, $\Sigma_2^{-1}\Sigma_1 A = A\Sigma_2^{-1}(m)\Sigma_1(m)$.

Obviously the criterion function *J* is invariant with respect to non-singular transforms and, using standard arguments, one can prove that $\Phi^{(m)} = (\Phi_1, ..., \Phi_m)$ can be taken as the suboptimal linear feature extractor where $\Phi_i, i = 1, m$ are unit eigenvectors corresponding to the eigenvalues $\lambda_1, ..., \lambda_m$ of $\Sigma_2^{-1}\Sigma_1$ such that

$\lambda_1 + \frac{1}{\lambda_1} \geq ... \geq \lambda_m + \frac{1}{\lambda_m} \geq ... \geq \lambda_n + \frac{1}{\lambda_n}$.

But, in case of image restoration problem, both assumptions $\mu_1 = \mu_2$, $\Sigma_1 = \Sigma_2$ are unrealistic, therefore, we are forced to accept the hypothesis that $\mu_1 \neq \mu_2$ and $\Sigma_1 \neq \Sigma_2$. Since there is no known procedure available to optimize the criterion *J* when $\Sigma_1 \neq \Sigma_2$ and $\mu_1 \neq \mu_2$, a series of attempts to find suboptimal feature extractors have been looked for, as for instance [15].

5 Concluding remarks

The paper presents a series of developments aiming to obtain a suitable methodology for image denoising and restoration purposes in a model-free framework. The performances of the resulted algorithms were experimentally evaluated by a long series of tests entailing positive and optimistic conclusions. Some of the presented results and variants of them have been already published and they were welcomed by the scientific community.

A series of new developments extending this work are in progress and the experimental analysis points out a promising perspective.

References

[1] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, 1959

[2] S. Bacchelli and S. Papi, "Image denoising using principal component analysis in the wavelet domain", *Journal of Computational and Applied Mathematics*, vol. 189, no. 1-2, 1 May 2006.

[3] E. J. Balster, Y. F. Zheng and R. L. Ewing, "Fast, Feature-Based Wavelet Shrinkage Algorithm for Image Denoising. In: International Conference on Integration of Knowledge IntensiveMulti-Agent Systems". *KIMAS '03: Modeling, Exploration, and Engineering Held* Cambridge, MA on 30 September – 4 October, 2003.

[4] S. Beheshti and M.A. Dahleh, "A new information-theoretic approach to signal denoising and best basis selection" In: *IEEE Trans. On Signal Processing*, vol. 53, no. 10, pp. 3613- 3624, 2003.

[5] A. Buades, B. Coll and J.-M. Morel, "A non-local algorithm for image denoising", In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. vol. 2, pp: 60 – 65, 20-25 June 2005

[6] C. Chatterjee, V.P. Roychowdhury and E.K.P. Chong, "On Relative Convergence Properties of Principal Component Analysis Algorithms". *IEEE Transaction on Neural Networks*, vol. 9, no. 2, 1998.

[7] R. Chellappa and H. Jinchi, "A nonrecursive Filter for Edge Preserving Image Restoration" In: *Proc. Intl.Conf.on Acoustic, Speech and Signal Processing*, Tampa 1985

[8] C. Cocianu, L. State, D. Constantin and C. Săraru, "Partially Supervised Approach in Signal Recognition", *Informatică Economică*, vol.13, no.3, pp.153-164, 2009

[9] C. Cocianu, L. State and P. Vlamos, "Principal Axes – Based Classification with Application in Gaussian Distributed Sample Recognition", Economic *Computation and Economic Cybernetics Studies and Research*, vol. 41, no 1-2, pp. 159-166, 2007.

[10] C. Cocianu, L. State, V. Ştefănescu, P. Vlamos, "On the Efficiency of a Certain Class of Noise Removal Algorithms in Solving Image Processing Tasks", In: *Proceedings of the ICINCO*, Setubal, Portugal, 2004.

[11] C. Cocianu, L. State and P. Vlamos, "On a Certain Class of Algorithms for Noise Removal in Image Processing:A Comparative Study", In: *Proceedings of the Third IEEE Conference on Information Technology ITCC-2002*, Las Vegas, Nevada, USA, April 8-10, 2002

[12] G. Deng, D. Tay and S. Marusic, "A signal denoising algorithm based on overcomplete wavelet representations and Gaussian models". *Signal Processing* vol. 87, no. 5, pp. 866-876, May 2007.

[13] K. I. Diamantaras and S.Y. Kung, *Principal Component Neural Networks: theory and applications*. John Wiley &Sons

[14] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley&Sons, 1973

[15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press,Inc. 1990

[16] R. Gonzales and R. Woods, "*Digital Image Processing*", Prentice Hall

[17] S. Haykin, *Neural Networks A Comprehensive Foundation*, Prentice Hall, Inc.

[18] A. Hyvarinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley &Sons

[19] A. Hyvarinen, P. Hoyer, P. Oja, (1999) Image Denoising by Sparse Code Shrinkage, www.cis.hut.fi/projects/ica,

[20] A. K. Jain, R. Kasturi and B. G. Schnuck, *Machine Vision*, McGraw Hill

[21] J. Karhunen, E. Oja, "New Methods for Stochastic Approximations of Truncated Karhunen-Loeve Expansions", *Proceedings 6th International Conference on Pattern Recognition*, Springer Verlag, 1982.

[22] I. Pitas, *Digital Image Processing Algorithms*, Prentice Hall

[23] O. Rioul and M. Vetterli, "Wavelets and signal processing". *IEEE Signal Processing* Mag. vol. n, no. 8, pp. 14-38, 1991.

[24] I. Roşca, L. State, C. Cocianu, "Learning Schemes in Using PCA Neural Networks for Image Restoration Purposes". *WSEAS Transactions on Information Science and Applications*, vol. 5, pp. 1149-1159, 2008.

[25] T.D. Sanger, *An Optimality Principle for Unsupervised Learning, Advances in Neural Information Systems*, ed. D.S. Touretzky, Morgan Kaufmann

[26] M. Sonka and V. Hlavac, *Image Processing, Analyses and Machine Vision*, Chapman & Hall Computing

[27] J.L. Stark, F. Murtagh, A. Bijaoui, *Multiresolution Support Applied to Image Filtering and Restoration*, Technical Report, 1995.

[28] L. State, I. Paraschiv-Munteanu, *Introducere in teoria statistica a recunoasterii formelor*. Editura Universitatii din Pitesti, 2009.

[29] L. State, C. Cocianu, V. Panayiotis, "Attempts in Using Statistical Tools for Image Restoration Purposes", *The Proceedings of SCI2001*, Orlando, USA, July 22-25, 2001.

[30] L. State, C. Cocianu, C. Săraru, P. Vlamos, "New Approaches in Image Compression and Noise Removal", *Proceedings of the First International Conference on Advances in Satellite and Space Communications, SPACOMM*, IEEE Computer Society Press, 2009

[31] L. State, C. Cocianu, P. Vlamos, "A New Unsupervized Learning Scheme to Classify Data of Relative Small Volume". *Economic Computation and Economic Cybernetics Studies and Research*, pp. 109-120, 2008.

[32] L. State, C. Cocianu and P. Vlamos, "The Use of Features Extracted from Noisy Samples for Iimage Restoration Purposes", *Informatica Economică*, no. 41, 2007.

[33] L. State, C. Cocianu, P. Vlamos and V. Ştefănescu, "PCA-Based Data Mining Probabilistic and Fuzzy Approaches with Applications in Pattern Recognition".In: *Proceedings of ICSOFT*, 2006.

[34] L. State, C. Cocianu, P. Vlamos, V. Ştefănescu, "Noise Removal Algorithm Based on Code Shrinkage Technique", *Proceedings of the 9th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, USA, 2005.

[35] S. Umbaugh, *Computer Vision and Image Processing*, Prentice Hall, 1998.

Luminiţa STATE is professor, PhD., currently working with University of Pitesti, Department of Mathematics and Computer Science. Competence areas: artificial intelligence, machine learning, statistical pattern recognition, digital image processing. Research in the fields of machine learning, pattern recognition, neural computation. Author of 15 books and more than 120 papers published in national and international journals.

Cătălina-Lucia COCIANU is professor, PhD., currently working with Academy of Economic Studies, Faculty of Cybernetics, Statistics and Informatics, Department of Informatics in Economy. Competence areas: statistical pattern recognition, digital image processing. Research in the fields of pattern recognition, data mining, signal processing. Author of 12 books and more than 80 papers published in national and international journals.