Chapter 7

Application of Principal Component Analysis to Image Compression

Wilmar Hernandez and Alfredo Mendez

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75007

Abstract

In this chapter, an introduction to the basics of principal component analysis (PCA) is given, aimed at presenting PCA applications to image compression. Here, concepts of linear algebra used in PCA are introduced, and PCA theoretical foundations are explained in connection with those concepts. Next, an image is compressed by using different principal components, and concepts such as image dimension reduction and image reconstruction quality are explained. Also, using the almost periodicity of the first principal component, a quality comparative analysis of a compressed image using two and eight principal components is carried out. Finally, a novel construction of principal components by periodicity of principal components has been included, in order to reduce the computational cost for their calculation, although decreasing the accuracy.

Keywords: principal component analysis, population principal components, sample principal components, image compression, image dimension reduction, image reconstruction quality

1. Introduction

Principal component analysis, also known as the Hotelling transform or Karhunen-Loeve transform, is a statistical technique that was proposed by Karl Pearson (1901) as part of factorial analysis; however, its first theoretical development appeared in 1933 in a paper written by Hotelling [1–8]. The complexity of the calculations involved in this technique delayed its development until the birth of computers, and its effective use started in the second half of the twentieth century. The relatively recent development of methods based on principal components makes them little used by a large number of non-statistician researchers. The purposes of these

notes are to disclose the nature of the principal component analysis and show some of its possible applications.

Principal component analysis refers to the explanation of the structure of variances and covariances through a few linear combinations of the original variables, without losing a significant part of the original information. In other words, it is about finding a new set of orthogonal axes in which the variance of the data is maximum. Its objectives are to reduce the dimensionality of the problem and, once the transformation has been carried out, to facilitate its interpretation.

By having p variables collected on the units analyzed, all are required to reproduce the total variability of the system, and sometimes the majority of this variability can be found in a small number, k, of principal components. Its origin lies in the redundancy that there exists many times between different variables, so the redundancy is data, not information. The k principal components can replace the p initial variables, so that the original set of data, consisting of n measures of p variables, is reduced to n measures of k principal components.

The objective pursued by the analysis of principal components is the representation of the numerical measurements of several variables in a space of few dimensions, where our senses can perceive relationships that would otherwise remain hidden in higher dimensions. The abovementioned representation must be such that, when discarding higher dimensions, the loss of information is minimal. A simile could illustrate the idea: imagine a large rectangular plate that is a three-dimensional object, but that for practical purposes, we consider it as a flat two-dimensional object. When carrying out this reduction in dimensionality, a certain amount of information is lost since, for example, opposite points located on the two sides of the rectangular plate will appear confused in a single one. However, the loss of information is largely compensated by the simplification made, since many relationships, such as the neighborhood between points, are more evident when they are drawn on a plane than when done by a three-dimensional figure that must necessarily be drawn in perspective.

The analysis of principal components can reveal relationships between variables that are not evident at the first sight, which facilitates the analysis of the dispersion of observations, highlighting possible groupings and detecting the variables that are responsible for the dispersion.

2. Preliminaries

The study of multivariate methods is greatly facilitated by means of matrix algebra [9–11]. Next, we introduce some basic concepts that are essential for the explanation of statistical techniques, as well as for geometric interpretations. In addition, the relationships that can be expressed in terms of matrices are easily programmable on computers, so we can apply calculation routines to obtain other quantities of interest. It is a basic introduction about concepts and relationships.

2.1. The vector of means and the covariance matrix

Let $\mathbf{X} = \begin{bmatrix} X_1 & \dots & X_p \end{bmatrix}^t$ be a random column vector of dimension p. Each component, X_i, is a random variable (r.v.) with mean $E[X_i] = \mu_i$ and variance $V[X_i] = E\left[\left(X_i - \mu_i\right)^2\right] = \sigma_{ii}$. Given

two r.v., X_i and X_j, we define the covariance between them as $Cov[X_i, X_j] = E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ij}$. The expected values, variances, and covariances can be grouped into vectors and matrices that we will call population mean vector, $\boldsymbol{\mu}$, and population covariance matrix, $\sum$:

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \sum = Cov[\mathbf{X}] = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^t\right] = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix} \quad (1)$$

The population correlation matrix is given by $\boldsymbol{\rho} = \left[\rho_{ij}\right]$, where $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$.

In the case of having n values of the r.v.s, we will consider estimators of the previous population quantities, which we will call sample estimators.

Definition 2.1: *Let* $\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pp} \end{bmatrix}$ *be a simple random sample of a p-dimensional r.v. ordered in the data matrix, with the values of the r.v.s in each column. The p-dimensional sample mean column vector is* $\overline{\mathbf{X}} = [\overline{x}_i]$, *where* $\overline{x}_i = \frac{1}{p}\sum_{m=1}^{p} x_{im}$. *The sample covariance matrix is* $\mathbf{S} = \left[s_{ij}\right] = \frac{n}{n-1}\mathbf{S_n} = \frac{n}{n-1}\left(\mathbf{X}\text{-}\overline{\mathbf{X}}\right)\left(\mathbf{X}\text{-}\overline{\mathbf{X}}\right)^t$. *The generalized sample variance is the determinant of* $\mathbf{S}$, $|\mathbf{S}|$. *The sample correlation matrix is* $\mathbf{R} = \left[r_{ij}\right]$, *where* $r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$ *with* $i, j = 1 \ldots p$.

Proposition 2.1: *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_p$ *be a simple random sample of a p-dimensional r.v.* $\mathbf{X}$ *with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\sum$. *The unbiased estimators of* $\boldsymbol{\mu}$ *and* $\sum$ *are* $\overline{\mathbf{X}}$ *and* $\mathbf{S}$..

2.2. Eigenvalues and eigenvectors

One of the problems that linear algebra deals with is the simplification of matrices through methods that produce diagonal or triangular matrices, which are widely used in the resolution of linear systems of the form $\mathbf{Ax} = \mathbf{b}$.

Definition 2.2: *Let* $\mathbf{A}$ *be a square matrix. If* $\mathbf{v^t Av} \geq 0$ *for any vector* $\mathbf{v}$, $\mathbf{A}$ *is a nonnegative definite matrix. If* $\mathbf{Av} = \lambda\mathbf{v}$, *with* $\mathbf{v} \neq 0$, λ *is an eigenvalue associated with the eigenvector* $\mathbf{v}$.

Proposition 2.2: *Let* $\mathbf{A}$ *be a symmetric p by p matrix with real-valued entries.* $\mathbf{A}$ *has p pairs of eigenvalues and eigenvectors,* $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$, *such that:*

1. *All the eigenvalues are real. Also,*
   - **a.** $\mathbf{A}$ *is positive definite if all the eigenvalues are positive.*
   - **b.** $\mathbf{A}$ *is nonnegative definite if all the eigenvalues are nonnegative.*

2. *The eigenvectors can be chosen with 2-norm equal to 1.*
3. *The eigenvectors are mutually perpendicular.*
4. *The eigenvectors are unique unless two or more eigenvalues are equal.*
5. *The spectral decomposition of* $\mathbf{A}$ *is* $\mathbf{A} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^t + \cdots + \lambda_p \mathbf{e}_p \mathbf{e}_p^t$.
6. *If* $\mathbf{P} = \left[\mathbf{e}_1, \ldots, \mathbf{e}_p\right]$ *is an orthogonal matrix and* Λ *is a diagonal matrix with main diagonal entries* $\left(\lambda_1, \ldots, \lambda_p\right)$*, the spectral decomposition of* $\mathbf{A}$ *can be given by* $\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}^t$*. Therefore,* $\mathbf{A}^{-1} = \mathbf{P}\Lambda^{-1}\mathbf{P}^t = \sum_{i=1}^{p} \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i^t$.

Remark 2.1: *Let* $\mathbf{X}$ *be a matrix with the values of a simple random sample in each column of a p-dimensional r.v., and let* $\mathbf{y}_i^t = (x_{i1}, \ldots, x_{in})$*, with* $i = 1 \ldots p$*, be the ith row of* $\mathbf{X}$*. Let* $\mathbf{1}_n^t = (1, \ldots, 1)$ *be the n by one vector with all its coordinates equal to 1. It can be proven that:*

1. *The projection of the vector* $\mathbf{y}_i^t$ *on the vector* $\mathbf{1}_n$ *is the vector* $\overline{x}_i \mathbf{1}_n$*, whose 2-norm is equal to* $\sqrt{n}|\overline{x}_i|$.
2. *Matrix* $\mathbf{S_n}$ *is obtained from the residuals* $\mathbf{e}_i = \mathbf{y}_i - \overline{x}_i \mathbf{1}_n$*, the squared 2-norm of* $\mathbf{e}_i$ *is equal to* $(n-1)s_{ii}$*, and the scalar product of* $\mathbf{e}_i$ *and* $\mathbf{e}_j$ *is equal to* $(n-1)s_{ij}$.
3. *The sample correlation coefficient* r_{ij} *is the cosine of the angle between* $\mathbf{e}_i$ *and* $\mathbf{e}_j$.
4. *If* $\mathbf{U}$ *is the volume generated by the vectors* $\mathbf{e}_i$*, with* $i = 1 \ldots p$*, then* $|\mathbf{S}| = \frac{\mathbf{U}^2}{(n-1)^p}$*. Therefore, the generalized sample variance is proportional to the square of the volume generated by deviation vectors. The volume will increase if the norm of some* $\mathbf{e}_i$ *is increased.*

2.3. Distances

Many techniques of multivariate statistical analysis are based on the concept of distance. Let $Q = (x_1, x_2)$ be a point in the plane. The Euclidean distance from Q to the origin, O, is $d(Q, O) = \sqrt{x_1^2 + x_2^2}$. If $Q = \left(x_1, \ldots, x_p\right)$ and $R = \left(y_1, \ldots, y_p\right)$, the Euclidean distance between these two points of $\Re^p$ is $d(Q, R) = \sqrt{\left(x_1 - y_1\right)^2 + \cdots + \left(x_p - y_p\right)^2}$. All points $\left(x_1, \ldots, x_p\right)$ whose square distance to the origin is a fixed quantity, for example, $x_1^2 + \cdots + x_p^2 = c^2$, are the points of the p-dimensional sphere of radius $|c|$.

For many statistical purposes, the Euclidean distance is unsatisfactory, since each coordinate contributes in the same way to the calculation of such a distance. When the coordinates represent measures subject to random changes, it is desirable to assign weights to the coordinates depending on how high or low the variability of the measurements is. This suggests a measure of distance that is different from the Euclidean.

Next, we introduce a statistical distance that will take into account the different variabilities and correlations. Therefore, it will depend on the variances and covariances, and this distance is fundamental in multivariate analysis.

Suppose we have a fixed set of observations in $\Re^p$, and, to illustrate the situation, consider n pairs of measures of two variables, x_1 and x_2. Suppose that the measurements of x_1 vary independently of x_2 and that the variability of the measures of x_1 are much greater than those of x_2. This situation is shown in **Figure 1**, and our first objective is to define a distance from the points to the origin.

In **Figure 1**, we see that the values that have a given deviation from the origin are farther from the origin in the x_1 direction than in the x_2 direction, due to the greater variability inherent in the direction of x_1. Therefore, it seems reasonable to give more weight in the coordinate x_2 than in the x_1. One way to obtain these weights is to standardize the coordinates, that is, $x_1^* = x_1/\sqrt{s_{11}}$ and $x_2^* = x_2/\sqrt{s_{22}}$, where s_{ii} is the sample variance of the variable x_i. Thus, the statistical distance from a point $Q = (x_1, x_2)$ to the origin is $d(Q, O) = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$. Therefore, the points that are equidistant from the origin of a constant distance c are on an ellipse centered at the origin, whose major axis coincides with the coordinate that has the greatest variability. In the case that the variability of one variable is analogous to that of the other and that the coordinates are independent, the Euclidean distance is proportional to the statistical distance.

If $Q = (x_1, \ldots, x_p)$ and $R = \left(y_1, \ldots, y_p\right)$ are two points of $\Re^p$, the statistical distance between them is $d(Q, R) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}$, with s_{ii} being the sample variance of the variable x_i. The statistical distance defined so far does not include most of the important cases where the variables are not independent. **Figure 2** shows a situation where the pairs (x_1, x_2) seem to have an increasing trend, so the sample correlation coefficient will be positive. In **Figure 2**,

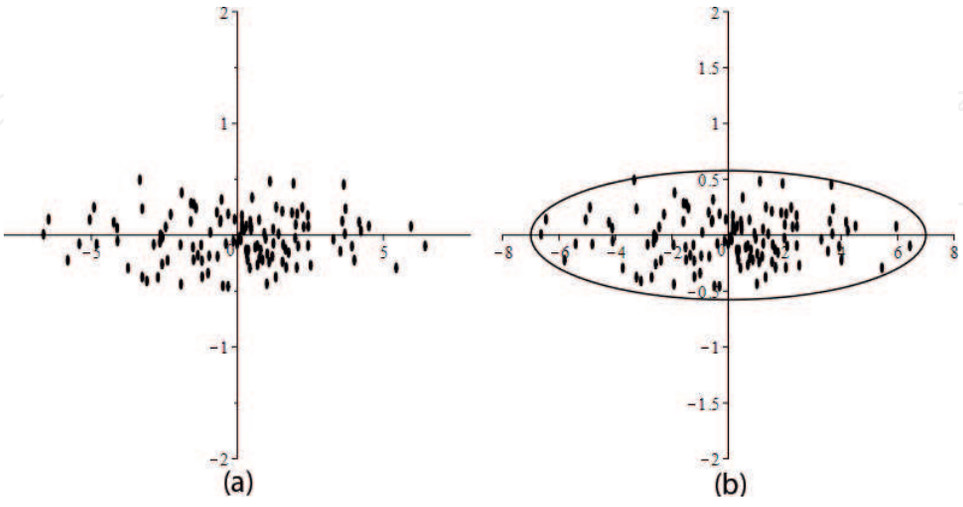

Figure 1. Scatter plot with more variability in x_1 than in x_2. (a)Scatter plot (b) Ellipse of constant distance.

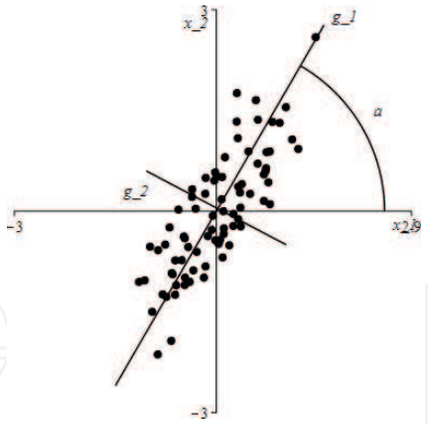

Figure 2. Scatter plot with positive correlation.

we see that if we make a rotation of amplitude α and consider the axes (g_1, g_2) we are in conditions analogous to those of **Figure 1 (a)**. Therefore, the distance from the point $Q = (g_1, g_2)$ to the origin will be $d(Q,O) = \sqrt{\frac{g_1^2}{\tilde{s}_{11}} + \frac{g_2^2}{\tilde{s}_{22}}}$, where $\tilde{s}_{ii}$ is the sample variance of the variable g_i.

The relationships between the original coordinates and the new coordinates can be expressed as

$$\begin{aligned} g_1 &= x_1\cos(\alpha) + x_2\sin(\alpha) \\ g_2 &= -x_1\sin(\alpha) + x_2\cos(\alpha) \end{aligned} \tag{2}$$

and, after some algebraic manipulations, $d(Q,O) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$, where a_{ij} are values that depend on the angle and the dispersions, and also must meet the condition that the distance between any two points must be positive.

The distance from a point $Q = (x_1, x_2)$ to a fixed point $R = (y_1, y_2)$ in situations where there is a positive correlation is $d(Q,R) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$. So, in this case, the coordinates of all points $Q = (x_1, x_2)$ verify the equation $a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2 = c^2$, which is the equation of an ellipse of center $R = (y_1, y_2)$ and with axes parallel to (g_1, g_2). **Figure 3** shows ellipses with constant statistical distances.

This distance can be generalized to $\mathfrak{R}^p$ if $a_{11}, \ldots, a_{pp}, a_{12}, \ldots, a_{p-1,p}$ are values such that the distance from Q to R is given by.

$$d(Q,R) = \sqrt{A + B}, \text{ where } \begin{aligned} A &= a_{11}(x_1 - y_1)^2 + \cdots + a_{pp}\left(x_p - y_p\right)^2 \\ B &= 2a_{12}(x_1 - y_1)(x_2 - y_2) + \cdots + 2a_{p-1,p}\left(x_{p-1} - y_{p-1}\right)\left(x_p - y_p\right) \end{aligned} \tag{3}$$

This distance, therefore, is completely determined by the coefficient a_{ij}, with $i, j \in \{1, \ldots, p\}$, which can be arranged in a matrix given by

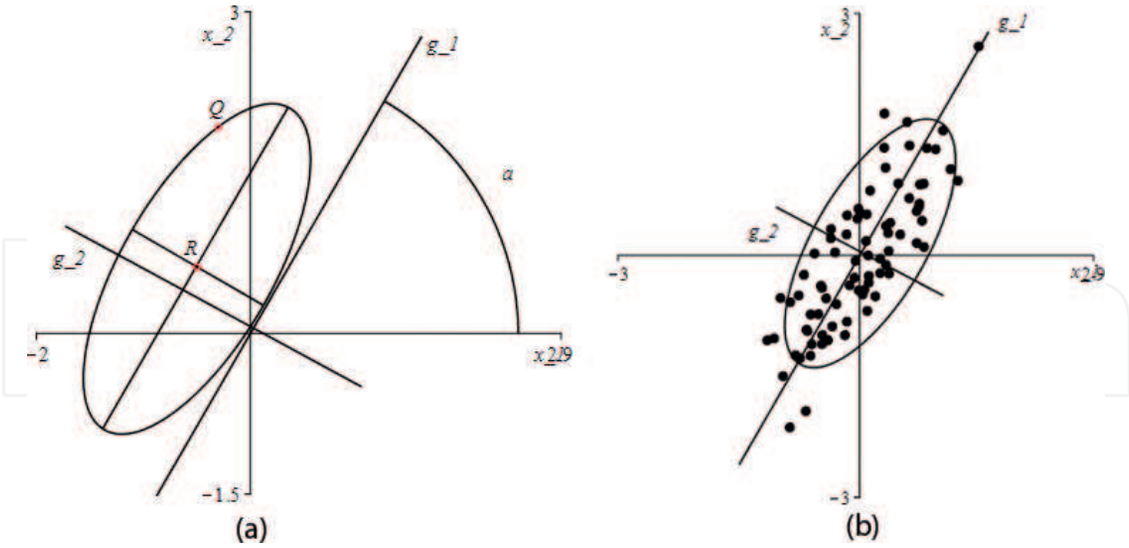

Figure 3. Ellipses of constant statistical distance. (a) Point Q at a constant distance from R. (b) Ellipse $x^2/3 + 4y^2 = 1$ rotated and moved and scatter plot.

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{bmatrix} \tag{4}$$

The elements of Eq. (4) cannot be arbitraries. In order to define a distance over a vector space, Eq. (4) must be a square, symmetric, positive definite matrix. Therefore, the sample covariance matrix of a data matrix, **S**, is a candidate to define a statistical distance.

Figure 4 shows a cloud of points with center of gravity, $(\overline{x}_1, \overline{x}_2)$, at point R. At the first glance, it can be seen that the Euclidean distance from point R to point Q is greater than the Euclidean

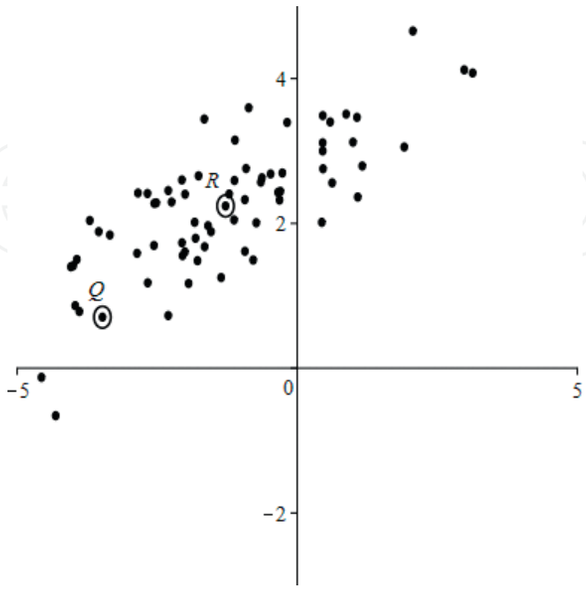

Figure 4. Scatter plot with center of gravity R and a point Q.

distance from point R to the origin; however, Q seems to have more to do with the cloud of points than the origin. If we take into account the variability of the points in the cloud and take the statistical measure, then Q will be closer to R than the origin.

The above given explanation has tried to be an illustration of the need to consider distances other than the Euclidean.

3. Population principal components

Principal components are a particular case of linear combinations of p r.v.s, $X_1, \ldots, X_p$. These linear combinations represent, geometrically, a new coordinate system that is obtained by rotating the original reference system that has $X_1, \ldots, X_p$ as coordinate axes. The new axes represent the directions with maximum variability and provide a simple description of the structure of the covariance.

Principal components depend only on the variance/covariance matrix $\sum$ (or on the correlation matrix $\boldsymbol{\rho}$) of $X_1, \ldots, X_p$, and it is not necessary to assume that the r.v.s follows an approximately normal distribution. In case of having a normal multivariate distribution, we will have interpretations in terms of ellipsoids of constant density, if we consider the distance that defines the $\sum$ matrix, and the inferences can be made from the population components.

Let $\mathbf{X} = \begin{bmatrix} X_1 & \ldots & X_p \end{bmatrix}^t$ be a p-dimensional random vector with covariance matrix $\sum$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Let us consider the following p linear combinations:

$$\begin{aligned} Y_1 &= l_1^t \mathbf{X} = l_{11} X_1 + \cdots + l_{p1} X_p \\ &\vdots \\ Y_p &= l_p^t \mathbf{X} = l_{1p} X_1 + \cdots + l_{pp} X_p \end{aligned} \tag{5}$$

These new r.v.s verify the following equalities:

$$\begin{aligned} &V[Y_i] = l_i^t \Sigma l_i \qquad i = 1, \ldots, p \\ &Cov\left[Y_i, Y_j\right] = l_i^t \Sigma l_j \quad i, j = 1, \ldots, p \quad i \neq j \end{aligned} \tag{6}$$

Principal components are those linear combinations that, being uncorrelated among them, have the greatest possible variance. Thus, the first principal component is the linear combination with the greatest variance, that is, $V[Y_1] = l_1^t \Sigma l_1$ is maximum. Since if we multiply l_1 by some constant the previous variance grows, we will restrict our attention to vectors of norm one, with which the aforementioned indeterminacy disappears. The second principal component is the linear combination that maximizes the variance and is uncorrelated with the first one, and the norm of the coefficient vector is equal to 1.

Proposition 3.1: *Let $\sum$ be the covariance matrix of the random vector $\mathbf{X} = \begin{bmatrix} X_1 & \ldots & X_p \end{bmatrix}^t$. Let us assume that $\sum$ has p pairs of eigenvalues and eigenvectors, $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Then, the ith principal component is given by*

http://dx.doi.org/10.5772/intechopen.75007

$$Y_i = \mathbf{e}_i^t \mathbf{X} = e_{1i}X_1 + \cdots + e_{pi}X_p \quad i = 1, \ldots, p \tag{7}$$

In addition, with this choice it is verified that:

1. $V[Y_i] = \mathbf{e}_i^t \Sigma \mathbf{e}_i = \lambda_i \quad i = 1, \ldots, p$.
2. $Cov\left[Y_i, Y_j\right] = 0 \quad i, j = 1, \ldots, p \quad i \neq j$.
3. *If any of the eigenvalues are equal, the choice of the corresponding eigenvectors as vectors of coefficients is not unique.*
4. $\sigma_{11} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} V[X_i] = \lambda_1 + \cdots + \lambda_p = \sum_{j=1}^{p} V\left[Y_j\right]$.

Remark 3.1: *For the demonstration of these results, expressions are used on maximums of quadratic forms between vectors of fixed norm* $\left(\max_{l \neq 0} \frac{l^t \Sigma l}{l^t l} = \lambda_1\right)$. *Also, the Lagrange multipliers method can be used, expressions when the abovementioned maximum is subject to orthogonality conditions and properties on the trace of a matrix (if* $\Sigma = \mathbf{P}\Lambda\mathbf{P}^{\mathrm{t}}$, *then* $tr(\Sigma) = tr\left(\mathbf{P}\Lambda\mathbf{P}^{\mathrm{t}}\right) = tr(\Lambda)$*).*

Due to the previous result, principal components are uncorrelated among them, with variances equal to the eigenvalues of $\sum$, *and the proportion of the population variance due to the ith principal component is given by* $\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}$.

If a high percentage of the population variance, for example, the 90%, of a p-dimensional r.v., with large p, can be attributed to, for example, the five first principal components, then we can replace all the r.v.s by those five components without a great loss of information.

Each component of the coefficient vector $\mathbf{e}_i^t = \left[e_{1i}, \ldots, e_{pi}\right]$, e_{ki}, also deserves our attention, since it is a measure of the relationship between the r.v.s $X_k$and Y_i.

Proposition 3.2: *If* $Y_1 = \mathbf{e}_1^t \mathbf{X}, \ldots, Y_p = \mathbf{e}_p^t \mathbf{X}$ *are the principal components obtained from the covariance matrix* $\sum$, *with pairs of eigenvalues and eigenvectors* $(\lambda_1, \mathbf{e}_1) \quad \ldots \quad \left(\lambda_p, \mathbf{e}_p\right)$, *then the linear correlation coefficients between the variables* X_k *and the components* Y_i *are given by*

$$\rho_{X_k, Y_i} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, \ldots, p \tag{8}$$

Therefore, e_{ki} is proportional to the correlation coefficient between $X_k$and Y_i.

In the particular case that $\mathbf{X}$ has a normal p-dimensional distribution, $\mathrm{N}_p\left(\mu, \Sigma\right)$, the density of $\mathbf{X}$ is constant in the ellipsoids with the center at $\boldsymbol{\mu}$ given by $(\mathbf{X} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = c^2$ that have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$ and $i = 1, \ldots, p$, where $(\lambda_i, \mathbf{e}_i)$ are the pairs of eigenvalues and eigenvectors of $\sum$.

If the covariance matrix, $\sum$, can be decomposed into $\Sigma = \mathbf{P}\Lambda\mathbf{P}^{\mathrm{t}}$, where $\mathbf{P}$ is orthogonal and Λ diagonal, it can be shown that $\Sigma^{-1} = \mathbf{P}\Lambda^{-1}\mathbf{P}^{\mathrm{t}} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i^t$. Also, if it can be assumed that $\boldsymbol{\mu} = 0$, to simplify the expressions, then

$$c^2 = \mathbf{x}^t \Sigma^{-1} \mathbf{x} = \frac{1}{\lambda_1} \left(\mathbf{e}_1^t \mathbf{x}\right)^2 + \frac{1}{\lambda_2} \left(\mathbf{e}_2^t \mathbf{x}\right)^2 + \cdots + \frac{1}{\lambda_p} \left(\mathbf{e}_p^t \mathbf{x}\right)^2 \tag{9}$$

If the principal components $y_1 = \mathbf{e}_1^t \mathbf{x}, \ldots, y_p = \mathbf{e}_p^t \mathbf{x}$ are considered, the equation of the constant density ellipsoid is given by

$$c^2 = \frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + \cdots + \frac{1}{\lambda_p} y_p^2 \tag{10}$$

Therefore, the axes of the ellipsoid have the directions of the principal components.

Example 3.1: *Let X_1, X_2, X_3 be the three-unidimensional r.v.s and $\mathbf{X} = [X_1, X_2 X_3]^t$, with covariance matrix*

$$\Sigma = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 8 & -3 \\ 0 & -3 & 2 \end{bmatrix} \tag{11}$$

It can be verified that the pairs of eigenvalues and eigenvectors are $\left(\lambda_1 = 9.243, \mathbf{e}_1^t = [0 \quad 0.924 \quad -0.383]\right)$, $\left(\lambda_2 = 2, \mathbf{e}_2^t = [1 \quad 0 \quad 0]\right)$, and $\left(\lambda_3 = 0.757, \mathbf{e}_3^t = [0 \quad 0.383 \quad 0.924]\right)$. Therefore, the principal components are the following:

$$\begin{aligned} Y_1 &= \mathbf{e}_1^t \mathbf{X} = 0.924 X_2 - 0.383 X_3 \\ Y_2 &= \mathbf{e}_2^t \mathbf{X} = X_1 \\ Y_3 &= \mathbf{e}_3^t \mathbf{X} = 0.383 X_2 + 0.924 X_3 \end{aligned} \tag{12}$$

The norm of all the eigenvectors is equal to 1, and, in addition, the variable X_1 is the second principal component, because X_1 is uncorrelated with the other two variables.

The results of **Proposition 3.1** *can be verified for this data, for example, $V[Y_1] = 9.243$ and $Cov[Y_1, Y_2] = 0$. Also, $\sum_{i=1}^{3} V[X_i] = 2 + 8 + 2 = 12 = 9.243 + 2 + 0.757 = \sum_{j=1}^{3} V\left[Y_j\right]$. Thus, the proportion of the total variance explained by the first component is $\lambda_1/12 = 77\%$, and the one explained by the first two is $(\lambda_1 + \lambda_2)/12 = 93.69\%$, so that the components Y_1 and Y_2 can replace the original variables with a small loss of information.*

The correlation coefficients between the principal components and the variables are the following:

$$\begin{aligned} &\rho_{X_1, Y_1} = 0 \quad \rho_{X_2, Y_1} = 0.993 \quad \rho_{X_3, Y_1} = -0.823 \\ &\rho_{X_1, Y_2} = 1 \quad \rho_{X_2, Y_2} = 0 \quad \rho_{X_3, Y_2} = 0 \\ &\rho_{X_1, Y_3} = 0 \quad \rho_{X_2, Y_3} = 0.118 \quad \rho_{X_3, Y_3} = 0.568 \end{aligned} \tag{13}$$

In view of these values, it can be concluded that $X_2$and X_3 individually are practically equally important with respect to the first principal component, although this is not the case with respect to the third

component. If, in addition, it is assumed that the distribution of $\mathbf{X}$ *is normal,* $N_3(\mu, \Sigma)$*, with a null mean vector, ellipsoids of constant density* $\mathbf{x}^t\Sigma^{-1}\mathbf{x} = c^2$ *can be considered. An ellipsoid of constant statistical distance and projections is shown in* ***Figure 5****.*

The ellipsoid with $c^2 = 8$ *has been represented in* ***Figure 5 (a)****, together with its axes and the ellipsoid projections on planes parallel to the coordinate axes. The aforementioned projections are ellipses of red, green, and blue colors that are reproduced in* ***Figure 5 (b)****. Also, in this figure, the black ellipse obtained by projecting the ellipsoid on the plane determined by the first two main components has been represented. The equation of this ellipse is* $\frac{y_1^2}{a^2} + \frac{y_2^2}{b^2} = 8$*, where* $a = \frac{c}{\sqrt{\eta_1}}$ *and* $b = \frac{c}{\sqrt{\eta_2}}$*, with* η_1 *and* η_2 *being the two smallest eigenvalues of* Σ^{-1}*, and the axes are determined by* Y_1 *and* Y_2*. As can be seen, the diameters of the ellipse determined by the first two components are larger than the others. Therefore, the area enclosed by this ellipse is the largest of all, indicating that it is the one that gathers the greatest variability.*

3.1. Principal components with respect to standardized variables

The principal components of the normalized variables $Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \ldots, Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$ can also be considered, which in matrix notation is $\mathbf{Z} = \mathbf{V}(\mathbf{X}\text{-}\boldsymbol{\mu})$, where $\mathbf{V}$ is the diagonal matrix whose elements are $\frac{1}{\sqrt{\sigma_{11}}}, \ldots, \frac{1}{\sqrt{\sigma_{pp}}}$. It is easily verified that the r.v. $\mathbf{Z}$ verifies $E[\mathbf{Z}] = 0$ and $Cov[\mathbf{Z}] = \mathbf{V}\Sigma\mathbf{V} = \boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is the correlation matrix of $\mathbf{X}$.

Principal components of $\mathbf{Z}$ are obtained by the eigenvalues and eigenvectors of the correlation matrix, $\boldsymbol{\rho}$, of $\mathbf{X}$. Furthermore, with some simplification, the previous results can be applied, since the variance of each Z_i is equal to 1.

Let $W_1, \ldots, W_p$ be the principal components of $\mathbf{Z}$ and $(v_i, \mathbf{u}_i^t)$, $i = 1, \ldots, p$, the pairs of eigenvalues and eigenvectors of $\boldsymbol{\rho}$, since they do not have to be the same.

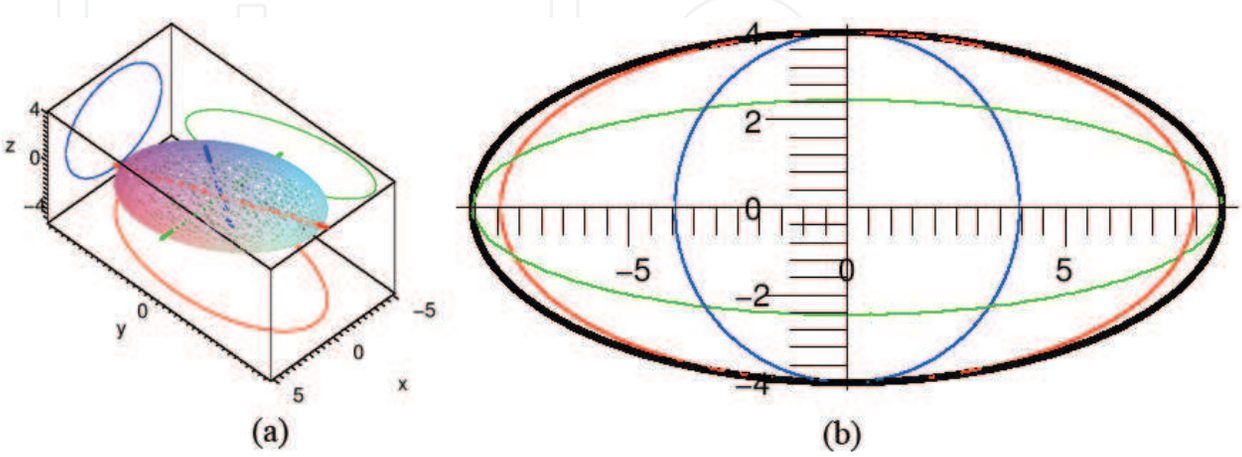

Figure 5. Ellipsoid of constant statistical distance and projections. (a) Ellipsoid of constant density and projections on the coordinate planes. (b) Projections on the coordinate planes and the base plane $\{Y_1, Y_2\}$.

Proposition 3.3: *Let* $\mathbf{Z} = [Z_1, \ldots, Z_p]^t$ *be a random vector with covariance matrix* $\boldsymbol{\rho}$. *Let* $(v_1, \mathbf{u}_1)$, …, $(v_p, \mathbf{u}_p)$ *be the pairs of eigenvalues and eigenvectors of* $\boldsymbol{\rho}$, *with* $v_1 \geq \cdots \geq v_p$. *Then, the ith principal component is given by* $W_i = \mathbf{u}_i^t \mathbf{V}(\mathbf{X}\text{-}\boldsymbol{\mu})$, $i = 1, \ldots, p$. *In addition, with this choice it is verified that:*

1. $V[W_i] = v_i$, $i = 1, \ldots, p$.
2. $Cov[W_i, W_j] = 0$, $i, j = 1, \ldots, p$, $i \neq j$.
3. *If any of the eigenvalues are equal, the choice of the corresponding eigenvectors as vectors of coefficients is not unique.*
4. $\sum_{i=1}^{p} V[W_i] = v_1 + \cdots + v_p = \sum_{j=1}^{p} V[Z_j] = p$.
5. *The linear correlation coefficients between the variables* Z_k *and the principal components* W_i *are* $\rho_{Z_k, W_i} = u_{ki}\sqrt{v_i}$ *and* $i, k = 1, \ldots, p$.

These results are a consequence of those obtained in **Proposition 3.1** and **Proposition 3.2** applied to **Z** and **ρ** instead of **X** and $\sum$.

The total population variance of the normalized variables is the sum of the elements of the diagonal of **ρ**, that is, p. Therefore, the proportion of the total variability explained by the ith principal component is $\frac{v_i}{p}$, $i = 1, \ldots, p$.

Example 3.2: *Let* X_1 *and* X_2 *be the two-unidimensional r.v.s and* $\mathbf{X} = [X_1, X_2]^t$ *with the covariance matrix,* $\sum$, *and correlation matrix,* **ρ**, *given by*

$$\Sigma = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$
$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} \tag{14}$$

It can be verified that the pairs of eigenvalues and eigenvectors for S are $(\lambda_1 = 100.04, \mathbf{e}_1^t = [-0.02 \quad -0.999])$ *and* $(\lambda_2 = 0.96, \mathbf{e}_2^t = [-0.999 \quad 0.02])$. *Therefore, the principal components are the following:*

$$Y_1 = \mathbf{e}_1^t \mathbf{X} = -0.02X_1 - 0.999X_2$$
$$Y_2 = \mathbf{e}_2^t \mathbf{X} = -0.999X_1 + 0.02X_2 \tag{15}$$

Furthermore, the eigenvalues and eigenvectors of **ρ** *are* $(v_1 = 1.2, \mathbf{u}_1^t = [0.707 \quad 0.707])$ *and* $(v_2 = 0.8, \mathbf{u}_2^t = [-0.707 \quad 0.707])$; *hence, the principal components of the normalized variables are the following:*

$$W_1 = \mathbf{u}_1^t \mathbf{Z} = 0.707Z_1 + 0.707Z_2 = 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2)$$
$$W_2 = \mathbf{u}_2^t \mathbf{Z} = -0.707Z_1 + 0.707Z_2 = -0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2) \quad (16)$$

Because the variance of X_2 is much greater than that of X_1, the first principal component for Σ is determined by X_2, and the proportion of variability explained by that first component is $\frac{\lambda_1}{\lambda_1+\lambda_2} = 0.99$.

When considering the normalized variables, each variable also contributes to the components determined by $\boldsymbol{\rho}$, and the dependencies between the normalized variables and their first component are $\rho_{Z_1,W_1} = u_{11}\sqrt{v_1} = 0.707\sqrt{1.2} = 0.774$ and $\rho_{Z_2,W_1} = u_{21}\sqrt{v_1} = -0.707\sqrt{1.2} = -0.774$. The proportion of the total variability explained by the first component is $\frac{v_1}{p} = 0.6$.

Therefore, the importance of the first component is strongly affected by normalization. In fact, the weights, in terms of X_i are 0.707 and 0.0707 for $\boldsymbol{\rho}$, as opposed to -0.02 and -0.999 for Σ.

Remark 3.2: *The above example shows that the principal components deduced from the original variables are, in general, different from those derived from the normalized variables. So, normalization has important consequences.*

When the units in which the different one-dimensional random variables are given are very different and in the case that one of the variances is very dominant compared to the others, the first principal component, with respect to the original variables, will be determined by the variable whose variance is the dominant one. On the other hand, if the variables are normalized, their relationship with the first components will be more balanced.

Principal components can be expressed in particular ways if the covariance matrix, or the correlation matrix, has special structures, such as diagonal ones, or structures of the form $\Sigma = \sigma^2\mathbf{A}$.

4. Sample principal components

Once we have the theoretical framework, we can now address the problem of summarizing the variation of n measurements made on p variables.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sample of a p-dimensional r.v. $\mathbf{X}$with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. These data have a vector of sample means $\overline{\mathbf{x}}$, covariance matrix $\mathbf{S}$, and correlation matrix $\mathbf{R}$.

This section is aimed at constructing linear uncorrelated combinations of the measured characteristics that contain the greatest amount of variability contained in the sample. These linear combinations are called principal sample components.

Given n values of any linear combination $l_1^t\mathbf{x}_j = l_{11}x_{1j} + \cdots + l_{p1}x_{pj}$, $j = 1, \ldots, n$, its sample mean is $l_1^t\overline{\mathbf{x}}_j$, and its sample variance is $l_1^t\mathbf{S}l_1$. If we consider two linear combinations, $l_1^t\mathbf{x}_j$ and $l_2^t\mathbf{x}_j$, their sample covariance is $l_1^t\mathbf{S}l_2$.

The first principal component will be the linear combination, $l_1^t \mathbf{x}_j$, which maximizes the sample variance, subject to the condition $l_1^t l_1 = 1$. The second component will be the linear combination, $l_2^t \mathbf{x}_j$, which maximizes the sample variance, subject to the condition that $l_2^t l_2 = 1$ and that the sample covariance of the pairs $\left(l_1^t \mathbf{x}_j, l_2^t \mathbf{x}_j\right)$ is equal to zero. This procedure is continued until the p principal components are completed.

Proposition 4.1: *Let* $\mathbf{S} = (s_{ik})$ *be the p by p matrix of sample covariances, whose pairs of eigenvalues and eigenvectors are* $\left(\hat{\lambda}_1, \hat{\mathbf{e}}_1\right), \ldots, \left(\hat{\lambda}_p, \hat{\mathbf{e}}_p\right)$, *with* $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. *Let* $\mathbf{x}$ *be an observation of the p-dimensional random variable* $\mathbf{X}$, *then:*

1. *The ith principal component is given by* $\hat{y}_i = \hat{\mathbf{e}}_i^t \mathbf{x} = \hat{e}_{1i} x_1 + \cdots + \hat{e}_{pi} x_p$, $i = 1, \ldots, p$.
2. *The sample variance of* $\hat{y}_k$ *is* $\hat{\lambda}_k$, $k = 1, \ldots, p$.
3. *The sample covariance of* $(\hat{y}_i, \hat{y}_k)$, $i \neq k$, *is equal to 0.*
4. *The total sample variance is* $\sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \cdots + \hat{\lambda}_p$.
5. *The sample correlation coefficients between* x_k *and* $\hat{y}_i$ *are* $r_{x_k, \hat{y}_i} = \frac{\hat{e}_{ki}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}$, $i, k = 1, \ldots, p$.

In the case that the random variables have a normal distribution, the principal components can be obtained from a maximum likelihood estimation $\hat{\Sigma} = \mathbf{S}_n$, and, in this case, the sampling principal components can be considered as maximum likelihood estimates of the population principal components. Although the eigenvalues of $\mathbf{S}$ and $\hat{\Sigma}$ are different but proportional, with constant proportionality fixed, the proportion of variability they explain is the same. The sample correlation matrix is the same for $\mathbf{S}$ and $\hat{\Sigma}$. We still do not consider the particular case of normal distribution of the variables, so as not to have to include hypotheses that should be verified for the data under study.

Sometimes, the observations $\mathbf{x}$ are centered by subtracting the mean $\overline{\mathbf{x}}$. This operation does not affect the covariance matrix and produces principal components of the form $\hat{y}_i = \hat{\mathbf{e}}_i^t(\mathbf{x} - \overline{\mathbf{x}})$, and in this case $\overline{\hat{y}}_i$ for any component, while the sample variances remain $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$.

When trying to interpret the principal components, the correlation coefficients $r_{x_k, \hat{y}_i}$ are more reliable guides than the coefficients $\hat{e}_{ik}$, since they avoid interpretive problems caused by the different scales in which the variables are measured.

4.1. Interpretations of the principal sample components

Principal sample components have several interpretations. If the distribution of $\mathbf{X}$ is close to $N_p(\boldsymbol{\mu}, \Sigma)$, then components $\hat{y}_i = \hat{\mathbf{e}}_i^t(\mathbf{x} - \overline{\mathbf{x}})$ are realizations of the main population components $Y_i = \mathbf{e}_i^t(\mathbf{X} - \boldsymbol{\mu})$, which will have distribution $N_p(\mathbf{0}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ is the diagonal matrix whose elements are the eigenvalues, ordered from major to minor, from the sample covariance

matrix. Keeping in mind the hypothesis of normality, contours of constant density, $E_p = \left\{\mathbf{x} \in \mathfrak{R}^p | (\mathbf{x} - \overline{\mathbf{x}})^t \mathbf{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) = c^2\right\}$, can be estimated and make inferences from them.

Although it is not possible to assume normality in the data, geometrically the data are n points $\mathfrak{R}^p$, and the principal components represent an orthogonal transformation whose coordinate axes are the axes of the ellipsoid E_p and with lengths proportional to $\sqrt{\hat{\lambda}_i}$, with $\hat{\lambda}_i$ being the eigenvalues of $\mathbf{S}$. Since all eigenvectors have been chosen such that their norm is equal to 1, the absolute value of the ith component $\left[\hat{y}_i = \left|\hat{\mathbf{e}}_i^t (\mathbf{x} - \overline{\mathbf{x}})\right|\right.$ is the length of the projection of the vector $(\mathbf{x} - \overline{\mathbf{x}})$ on the vector $\hat{\mathbf{e}}_i$. Therefore, the principal components can be seen as a translation of the origin to the point $\overline{\mathbf{x}}$ and a rotation of the axes until they pass through the directions with greater variability.

When there is a high positive correlation between all the variables and a principal component with all its coordinates of the same sign, this component can be considered as a weighted average of all the variables or the size of the index that forms that component. The components that have coordinates of different signs oppose a subset of variables against another, being a weighted average of two groups of variables.

The interpretation of the results is simplified assuming that the small coefficients are zero and rounding the rest to express the component as sums, differences, or quotients of variables.

The interpretation of the principal components can be facilitated by graphic representations in two dimensions. A usual graph is to represent two components as coordinate axes and project all points on those axes. These representations also help to test hypotheses of normality and to detect anomalous observations. If there is an observation that is atypical in the first variable, we will have that the variability in that first variable will grow and that the covariance with the other variables will decrease, in absolute value. Consequently, the first component will be strongly influenced by the first variable, distorting the analysis.

Sometimes, it is necessary to verify that the first components are approximately normal, although it is not reasonable to expect this result from a linear combination of variables that do not have to be normal.

The last component can help detect suspicious observations. Each observation $\mathbf{x}$ can be expressed as a linear combination of the eigenvectors of $\mathbf{S}$, $\mathbf{x}_j = \hat{y}_{1j}\hat{\mathbf{e}}_1 + \cdots + \hat{y}_{pj}\hat{\mathbf{e}}_p$, with which the difference between the first components $\hat{y}_{1j}\hat{\mathbf{e}}_1 + \cdots + \hat{y}_{qj}\hat{\mathbf{e}}_q$ and the observation $\mathbf{x}_j$ is $\hat{y}_{q-1j}\hat{\mathbf{e}}_{q-1} + \cdots + \hat{y}_{pj}\hat{\mathbf{e}}_p$, which is a vector with square of the norm $\hat{y}_{q-1j}^2 + \cdots + \hat{y}_{pj}^2$, and we will suspect of observations that have a large contribution to the square of the aforementioned norm.

An especially small value of the last eigenvalue of the covariance matrix, or correlation matrix, can indicate a linear dependence between the variables that have not been taken into account. In this case, some variable is redundant and should be removed from the analysis. If we have four variables and the fourth is the sum of the other three, then the last eigenvalue will be close to zero due to rounding errors, in which case we should suspect some dependence. In general, eigenvalues close to zero should not be ignored, and eigenvalues associated with these

eigenvalues can indicate linear dependencies in the data and cause deformations in the interpretations, calculations, and consequent analysis.

4.2. Standardized sample principal components

In general, principal components are not invariant against changes of scale in the original variables, as has been mentioned when referring to the normalized population principal components. Normalizing, or standardizing, the variables consists of performing the following transformation $\mathbf{z}_j = \mathbf{D}(\mathbf{x}_j - \overline{\mathbf{x}}) = \left[\frac{x_{1j}-\overline{x}_1}{\sqrt{s_{11}}}, \ldots, \frac{x_{pj}-\overline{x}_p}{\sqrt{s_{pp}}}\right]^t$, $j = 1, \ldots, p$. If the matrix $\mathbf{Z}$ is the p by n matrix whose columns are $\mathbf{z}_j$, it can be shown that its sample mean vector is the null vector and that its correlation matrix is the sample correlation matrix, $\mathbf{R}$, of the original variables.

Remark 4.1: *Applying that the principal components of the normalized variables are those obtained for the sample observations but substituting the matrix* $\mathbf{S}$ *for* $\mathbf{R}$*, we can establish that if* $\mathbf{z}_1, \ldots, \mathbf{z}_n$ *are the normalized observations, with covariance matrix* $\mathbf{R} = (r_{ik})$*, where* r_{ik} *is the sample correlation coefficient between observations* $\mathbf{x}_i$ *and* $\mathbf{x}_k$*, and if the pairs of eigenvalues and eigenvectors of* $\mathbf{R}$ *are* $(\hat{v}_1, \hat{\mathbf{u}}_1), \ldots, (\hat{v}_p, \hat{\mathbf{u}}_p)$*, with* $\hat{v}_1 \geq \cdots \geq \hat{v}_p \geq 0$*, then*

1. *The ith principal component is given by* $\hat{\omega}_i = \hat{\mathbf{u}}_i^t \mathbf{z} = \hat{u}_{1i} z_1 + \cdots + \hat{u}_{pi} z_p$*,* $i = 1, \ldots, p$*.*
2. *The sample variance of* $\hat{\omega}_k$ *is* $\hat{v}_k$*,* $k = 1, \ldots, p$*.*
3. *The sample covariance of* $(\hat{\omega}_i, \hat{\omega}_k)$*,* $i \neq k$*, is equal to 0.*
4. *The total sample variance is* $tr(\mathbf{R}) = p = \hat{v}_1 + \cdots + \hat{v}_p$*.*
5. *The sample correlation coefficients between* z_k *and* $\hat{\omega}_i$ *are* $r_{z_k, \hat{\omega}_i} = \hat{u}_{ki}\sqrt{\hat{v}_i}$*,* $i, k = 1, \ldots, p$*.*
6. *The proportion of the total sample variance explained by the ith principal component is* $\frac{\hat{v}_i}{p}$*.*

4.3. Criteria for reducing the dimension

The eigenvalues and eigenvectors of the covariance matrix, or correlation matrix, are the essence of the analysis of principal components, since the eigenvalues indicate the directions of maximum variability and the eigenvectors determine the variances. If a few eigenvalues are much larger than the rest, most of the variance can be explained with less than p variables.

In practice, decisions about the number of components to be considered must be made in terms of the pairs of eigenvalues and eigenvectors of the covariance matrix, or correlation matrix, and different rules have been suggested:

a. When performing the graph $\left(i, \hat{\lambda}_i\right)$, it has been empirically verified that with the first values there is a decrease with a linear tendency of quite steep slope and that from a certain eigenvalue this decrease is stabilized. That is, there is a point from which the eigenvalues are very similar. The criterion consists of staying with the components that exclude the small eigenvalues and that are approximately equal.

b. Select components until obtaining a proportion of the preset variance (e.g., 80%). This rule should be applied with care, since components that are interesting to reflect certain nuances suitable for the interpretation of the analysis could be excluded.

c. A rule that does not have a great theoretical support, which must be applied carefully so as not to discard any valid component for the analysis, but which has given good empirical results, is to retain those components with variances, $\hat{\lambda}_i$, above a certain threshold. If the work matrix is the correlation matrix, in which case the average value of the eigenvalues is one, the criterion is to keep the components associated with eigenvalues greater than unity and discard the rest.

5. Application to image compression

We are going to illustrate the use of principal components to compress images. To this end, the image of Lena was considered. This photograph has been used by engineers, researchers, and students for experiments related to image processing.

5.1. Black and white photography

The black and white photograph shown in **Figure 6** was considered. First, the image in .jpg format was converted into the numerical matrix **Image** of dimension 512 by 512 (i.e., 2^9x2^9). Second, to obtain the observation vectors, the matrix was divided into blocks of dimension 2^3x2^3, $\mathbf{A}_{ij}$, with which 4096 blocks were obtained, and each of them was a vector of observations.

Figure 6. Black and white photograph of Lena.

$$\mathbf{Image} = \begin{bmatrix} \mathbf{A}_{1,1} & \cdots & \mathbf{A}_{1,64} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{64,1} & \cdots & \mathbf{A}_{64,64} \end{bmatrix} \tag{17}$$

Third, each matrix $\mathbf{A}_{ij}$ was stored in a vector of dimension 64, $\mathbf{x}$, which contained the elements of the matrix by rows, that is, $\mathbf{x} = [a_{i,1}, \ldots, a_{i,8}, a_{i+1,1}, \ldots, a_{i+1,8}, \ldots, a_{i+8,8}]$. This way, we had the observations $\{\mathbf{x}_k \in \Re^{64} | k = 1, \ldots, 4096\}$, which were grouped in the observation matrix $\mathbf{x} = [x_{ij}] \in \mathrm{M}_{4096,64}(\Re)$.

Fourth, the average of each column, $\overline{\mathbf{x}} = [\overline{x}_1, \ldots, \overline{x}_{64}]$, was calculated obtaining the vector of means, and from each observation x_{ij}, its corresponding mean $\overline{x}_j$ was subtracted. Thus, the matrix of centered observations $\mathbf{U}$ was obtained. The covariance matrix of $\mathbf{x}$ was $\mathbf{S} = \mathbf{U}^t\mathbf{U} \in \mathrm{M}_{64,64}(\Re)$.

Fifth, the 64 pairs of eigenvalues and eigenvectors of $\mathbf{S}$, $\left(\hat{\lambda}_i, \hat{\mathbf{e}}_i\right)$, were found, and they were ordered according to the eigenvalues from highest to lowest. The 8 largest eigenvalues are drawn in **Figure 7**. As can be seen, the first eigenvalue is much larger than the rest. Thus, the first principal component completely dominates the total variability.

Sixth, with the theoretical results and the calculations previously made, the 64 principal components $\hat{y}_j = \hat{\mathbf{e}}_j^t\mathbf{x} = \hat{e}_{1,j}x_1 + \cdots + \hat{e}_{64,j}x_p$, $j = 1, \ldots, p$, were built. The first principal component was $\hat{y}_1 = -0.1167x_1 + \cdots - 0.1166x_{64}$. Therefore, an orthonormal basis of $\Re^{64}$ was built.

Seventh, each vector $\hat{\mathbf{e}}_j = \left[\hat{e}_{1,j}, \ldots, \hat{e}_{64,j}\right]^t$ was grouped by rows in a matrix $\mathrm{M}_{8,8}$:

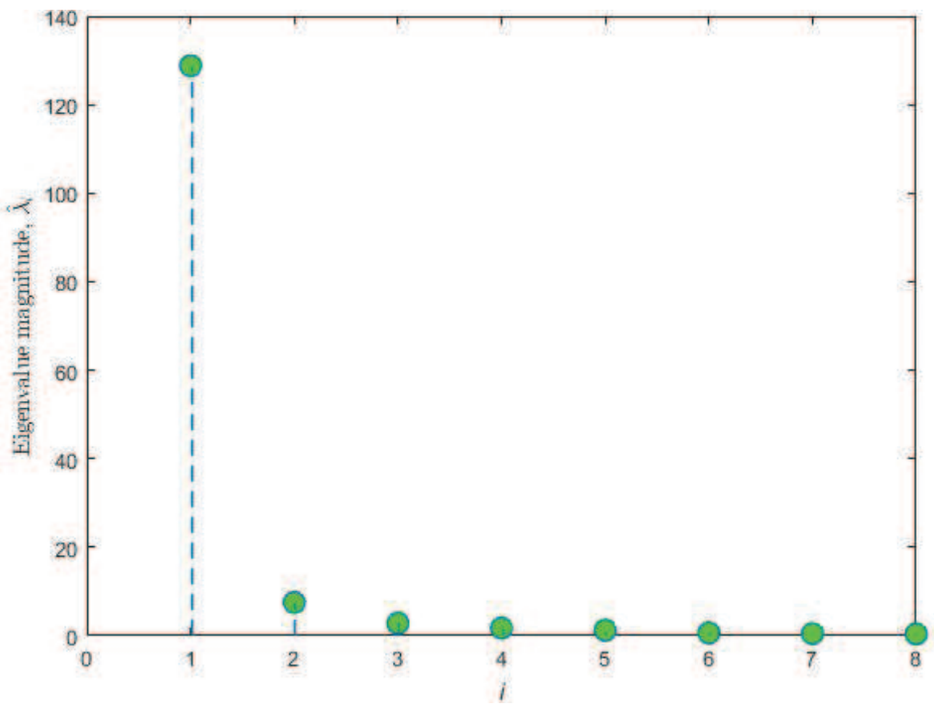

Figure 7. Graph $\left(i, \widehat{\lambda}_i\right)$, $i = 1, \ldots, 8$, with $\widehat{\lambda}_i$ being the eigenvalues ordered from highest to lowest.

$$\hat{\mathbf{E}}_j = \begin{bmatrix} \hat{e}_{1,j} & \cdots & \hat{e}_{8,j} \\ \vdots & \ddots & \vdots \\ \hat{e}_{57,j} & \cdots & \hat{e}_{64,j} \end{bmatrix} \tag{18}$$

Each of the 64 matrices $\hat{\mathbf{E}}_j$ was converted into an image. The images of the first three principal components are shown in **Figure 8**.

At this point, it is important to mention that the data matrix **x** has been assumed to be formed by 4096 vectors of $\mathcal{R}^{64}$ expressed in the canonical base, $\boldsymbol{B}$. Also, the base whose vectors were the eigenvectors of **S**, $\boldsymbol{B}' = \{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_{64}\}$, was considered. The coordinates with respect to the canonical basis of the vectors of $\boldsymbol{B}'$ were the columns of the matrix $\mathbf{PC} = \left[\hat{\mathbf{e}}_1^t, \ldots, \hat{\mathbf{e}}_{64}^t\right]$. Then, given a vector **v** that with respect to the canonical base had coordinates $(x_1, \ldots, x_{64})$ and with respect to the base $\boldsymbol{B}'$ had coordinates $(y_1, \ldots, y_{64})$, the relation between them was $[x_1, \ldots, x_{64}]^t = \mathbf{PC}[y_1, \ldots, y_{64}]^t$. Also, as **PC** is an orthogonal matrix, $[y_1, \ldots, y_{64}] = [x_1, \ldots, x_{64}]\mathbf{PC}$. Thus, the coordinates of the 4096 vectors that formed the observations matrix had as coordinates, with respect to the new base, the rows of the matrix of dimension 4096x64 given by $\mathbf{y} = \mathbf{x} \cdot \mathbf{PC}$.

Eight, in order to reduce the dimension, it was taken into consideration that if we keep all the vectors of $\boldsymbol{B}'$, we can perfectly reconstruct our data matrix, because $\mathbf{y} = \mathbf{x} \cdot \mathbf{PC} \Rightarrow \mathbf{x} = \mathbf{y} \cdot \mathbf{PC}^{-1} = \mathbf{y} \cdot \mathbf{PC}^t$. Additionally, for the case under study, to reduce the dimension, if we use the slope change rule, we can consider the first two principal components; five components if we want to explain 97% of the variability, because $\sum_{i=1}^{5} \hat{\lambda}_i / \sum_{j=1}^{64} \hat{\lambda}_j = 97\%$; or eight components if we want to explain 98% of the total variability.

In order to compress the image, the first vectors of the base $\boldsymbol{B}'$ were used. Moreover, supposing that we were left with M, $M < 64$, the matrix $\mathbf{T}_M$ given by Eq. (19) was defined:

$$\mathbf{T}_M = \begin{bmatrix} \mathbf{I}_{MxM} & \mathbf{0}_{Mx(64-M)} \\ \mathbf{0}_{(64-M)xM} & \mathbf{0}_{(64-M)x(64-M)} \end{bmatrix} \tag{19}$$

Therefore, the dimension of $\mathbf{y}_M = \mathbf{y} \cdot \mathbf{T}_M$ was 4096×64.

Ninth, to reconstruct the compressed image, each row of $\mathbf{y}_M$ was regrouped in an 8x8matrix. The ith row of $\mathbf{y}_M$, denoted by $\mathbf{y}_{Mi} = [b_{i,1}, \ldots, b_{i,8}, b_{i,9}, \ldots, b_{i,16}, \ldots, b_{i,64}]$, was transformed into

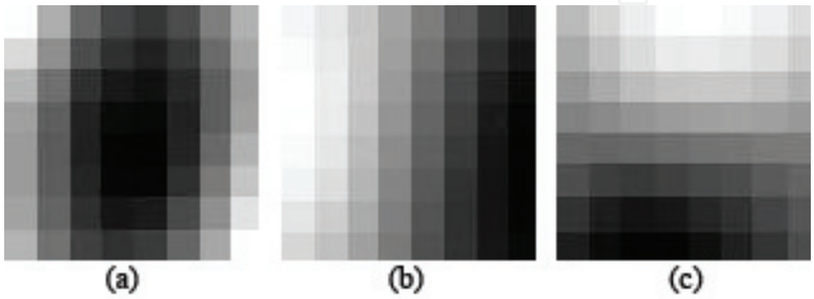

Figure 8. Images of the matrices of the first three principal components. (a) First component. (b) Second component. (c) Third component.

the matrix $\mathbf{B}_i$ given by Eq. (20), and the matrix **Compressed_image** given by Eq. (21) was built:

$$\mathbf{B}_i = \begin{bmatrix} b_{i,1} & \cdots & b_{i,8} \\ b_{i,9} & \cdots & b_{i,16} \\ \vdots & \ddots & \vdots \\ b_{i,57} & \cdots & b_{i,64} \end{bmatrix} \quad i = 1, \ldots, 4096 \tag{20}$$

$$\mathbf{Compressed\_image} = \begin{bmatrix} \mathbf{B}_1 & \cdots & \mathbf{B}_{64} \\ \mathbf{B}_{65} & \cdots & \mathbf{B}_{128} \\ \vdots & \ddots & \vdots \\ \mathbf{B}_{4033} & & \mathbf{B}_{4096} \end{bmatrix} \tag{21}$$

Figure 9. Original and compressed image with two, five, and eight principal components. (a) Original image. (b) Compression with two components. (c) Compression with five components. (d) Compression with eight components.

Tenth and finally, Eq. (21) was converted into a .jpg file. **Figure 9** shows the original image and compressed images with two, five, and eight principal components.

By increasing the number of principal components, the percentage of the variability explained is increased by very small percentages, but, nevertheless, nuances are added to the photo sufficiently remarkable, since they make it sharper, smooth out the contours, and mark the tones more precisely.

5.1.1. Objective measures of the quality of reconstructions

The two methods that we will use are the peak signal-to-noise ratio (*PSNR*) and the entropy of the error image. The PSNR measure evaluates the quality in terms of deviations between the processed and the original image, and the entropy of an image is a measure of the information content contained in that image.

Definition 5.1: *Let N be the number of rows by the number of columns in the image. Let $\{x_n | n = 1, \ldots, N\}$ be the set of pixels of the original image. Let $\{y_n | n = 1, \ldots, N\}$ be the set of reconstruction pixels. Let $\{r_n = x_n - y_n | n = 1, \ldots, N\}$ be the error. The mean square error (MSE) is*

$$MSE = \frac{1}{N}\sum_{n=1}^{N} r_n^2 \tag{22}$$

Definition 5.2: *Let the images under study be the 8 bit images. The peak signal-to-noise ratio of the reconstruction is*

$$PSNR = 10\log_{10}\left(\frac{\left(2^8 - 1\right)^2}{MSE}\right) \tag{23}$$

Figure 10 (a) shows *PSNR* of the reconstructions of the image versus the number of principal components used for the reconstruction, together with the regression line that adjusts the said cloud of points. **Figure 10 (b)** shows the values of the PSNR when we use three quarters (black), half (red), quarter (blue), eighth (green), sixteenth (brown), and the thirty-second part (yellow) of the components, which means a corresponding reduction in compression. A behavior close to linearity with a slope of approximately 0.2 can be seen. With the reductions considered, the *PSNR* varies between 27 and 63.

If the entropy is high, the variability of the pixels is very high, and there is little redundancy. Thus, if we exceed a certain threshold in compression, the original image cannot be recovered exactly. If the entropy is small, then the variability will be smaller. Therefore, the information of a pixel with respect to the pixels of its surroundings is high and, therefore, randomness is lost.

Definition 5.3: *Let I be an 8 bit image that can take the values $\{0, \ldots, 255\}$. Let p_i be the frequency with which the value $i \in \{0, \ldots, 255\}$ appears. Then, the entropy is*

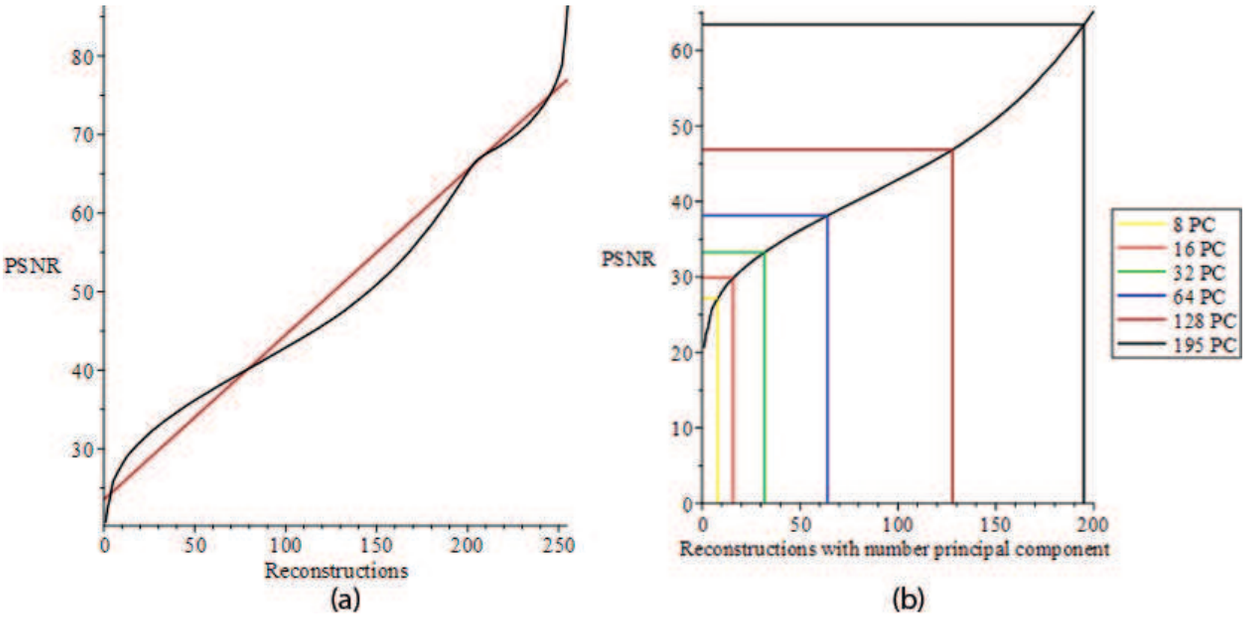

Figure 10. *PSNR* of the reconstructions according to the used principal components. (a) *PSNR*of 256 reconstructions. (b) *PSNR* of some reconstructions.

$$H(I) = -\sum_{i=0}^{255} p_i \log_2(p_i) \tag{24}$$

Figure 11 (a) shows the entropy of the reconstructions from 1 to 256 components. As can be seen, the entropy is increasing until the first 10 components, and then it becomes damped tending asymptotically to the value of the entropy of the image (7.4452). It can be seen that the difference with more than 170 components is insignificant. **Figure 11 (b)** shows the entropy of

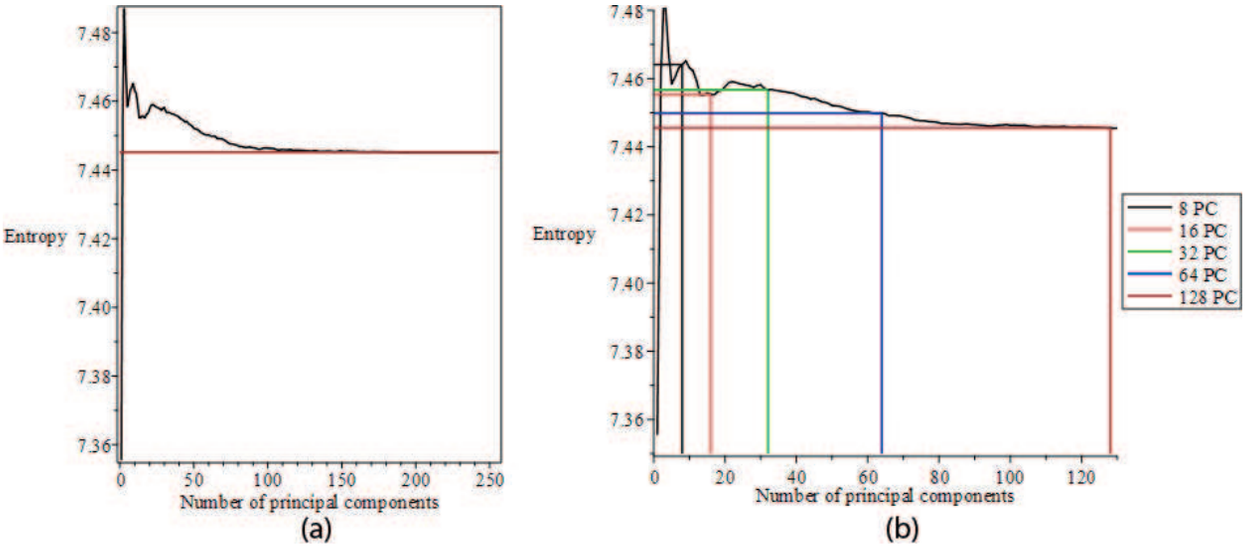

Figure 11. Entropy of reconstructions according to the used principal components. (a) Entropy of reconstructions. (b) Entropy of some reconstructions.

the reconstructions using 8 components (black), 16 components (brown), 32 components (green), 64 components (blue), and 128 components (red), respectively.

Finally, we consider the entropy of the images of the errors. Given an image, I, the value of each of its pixels is an element of the set $\{0, \ldots, 255\}$, and if we have a reconstruction, $\hat{I}$, and consider the error, $E = I - \hat{I}$, then the value of its pixels will be an element of the set $\{-255, \ldots, 255\}$. Therefore, E cannot be considered as an image. Since a pixel of value e_{ij} in E is an error of the same size as $-e_{ij}$, to consider images we denominate image of the error to $\mathrm{Im}(E) = \left[\left|e_{ij}\right|\right]$, being $E = \left[e_{ij}\right]$.

Figure 12 (a) shows the entropy of the error image versus the number of principal components used for the reconstruction, together with an adjusted line of slope -0.02. **Figure 12(b)** shows the entropy when we use 8 components (black), 16 components (brown), 32 components (green), 64 components (blue), and 128 components (red), respectively. With more than 200 principal components, the entropy of the errors is zero, which means that the errors have very little variability, and with fewer components, the decrease seems linear with slope -0.02.

5.2. Coordinates of the first principal component

In this section, we will consider the coordinates of the first vectors that form the principal components. If we consider that the vectors have been obtained as 2^3x2^3 dimension blocks, vectors will have 64 coordinates. **Figure 13** shows the coordinates of the first six principal components with respect to the canonical base.

As can be seen from **Figure 13**, all coordinates seem to have some component with period 8. This suggests that there may be some relationship with the shape of the blocks chosen and that most vectors are close to being periodic with period 8, because when we consider each of the

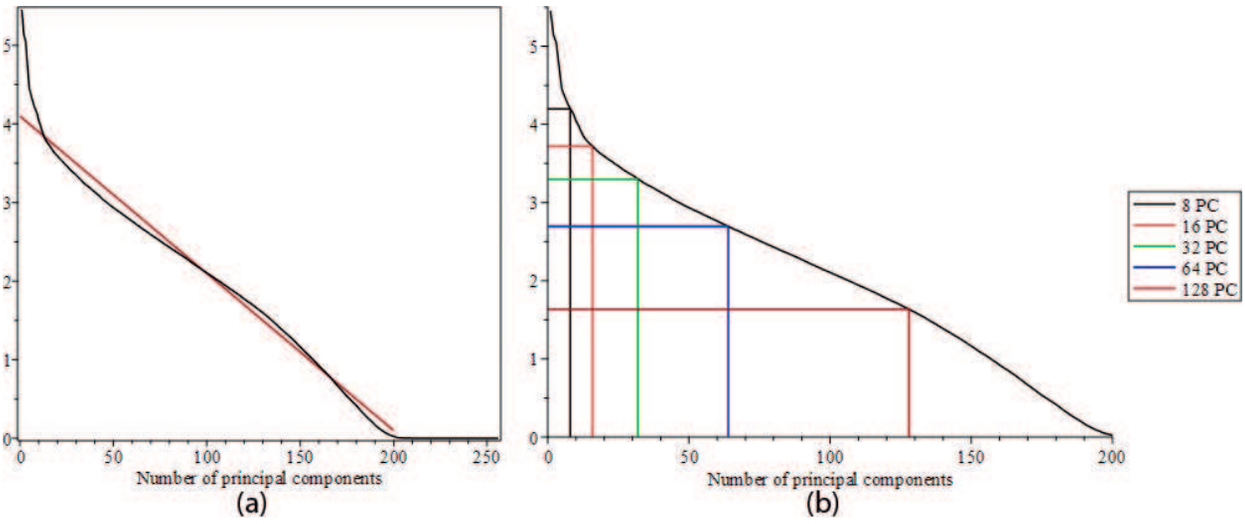

Figure 12. Entropy of the errors of the reconstructions converted into images according to the used principal components. (a) Entropy of differences (b) Entropy of some differences.

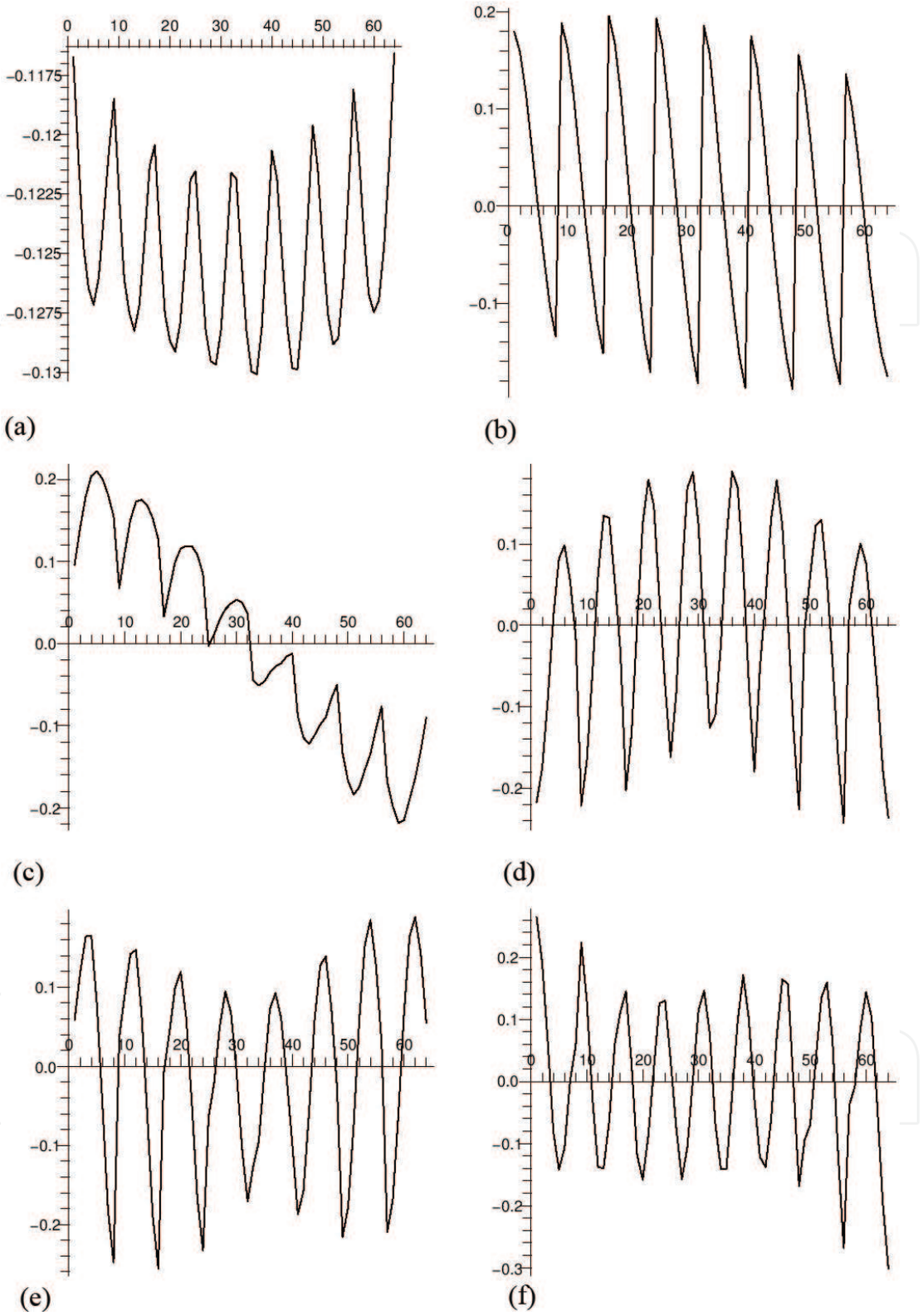

Figure 13. Coordinates of the first six principal components with respect to the canonical base. (a) First component. (b) Second component. (c) Third component. (d) Fourth component. (e) Fifth component. (f) Sixth component.

http://dx.doi.org/10.5772/intechopen.75007

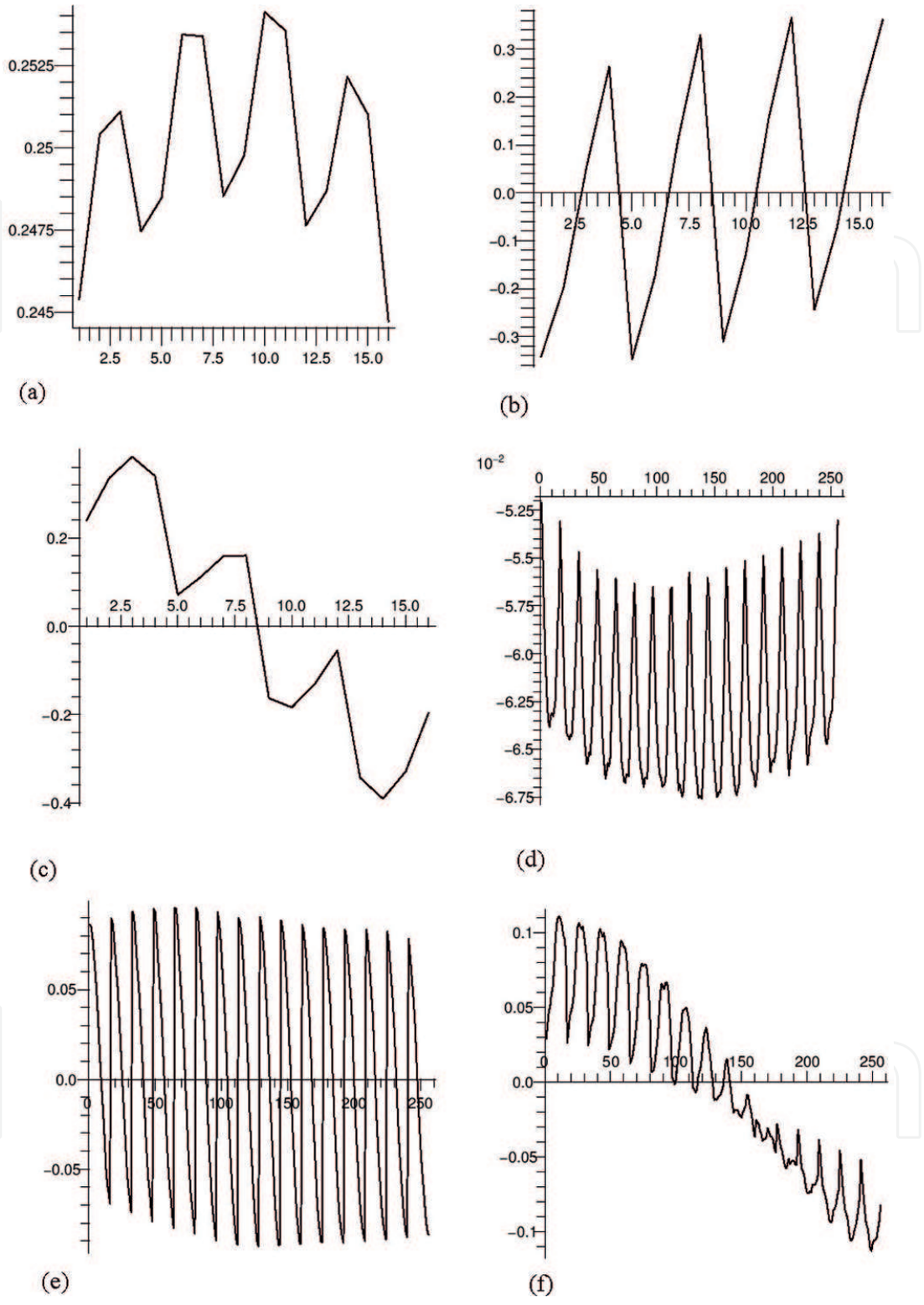

Figure 14. Coordinates of the first three principal components when vectors are constructed from blocks of 2^2x2^2 and 2^4x2^4. (a) First component 2^2x2^2 (b) Second component 2^2x2^2 (c) Third component 2^2x2^2 (d) First component 2^4x2^4 (e) Second component 2^4x2^4 (f) Third component 2^4x2^4.

4096 vectors of 64 components, the first 8 pixels are adjacent to the next 16 pixels, and these are adjacent to the next 8 pixels, and so on, up to 8 times.

Since the first principal components collect a large part of the characteristics of the vectors, it is plausible that they also reflect the periodicity of the vectors. Recall that principal components are linear combinations of vectors and that if all of them had all their periodic coordinates with the same period, then all components would be periodic as well.

In **Figure 14**, the coordinates of the first three principal components are shown when the vectors are constructed from blocks of 2^2x2^2(see **Figure 14 (a-c)**) and from blocks of 2^4x2^4 (see **Figure 14 (d-f)**). As can be seen, the periodicity in the first components is again appreciated.

Figure 15. Compression with 2 and 8 original and periodic principal components. (a) Compression with two components (b) Compression with two $components_{per}$. (c) Compression with eight components. (d) Compression with eight $components_{per}$.

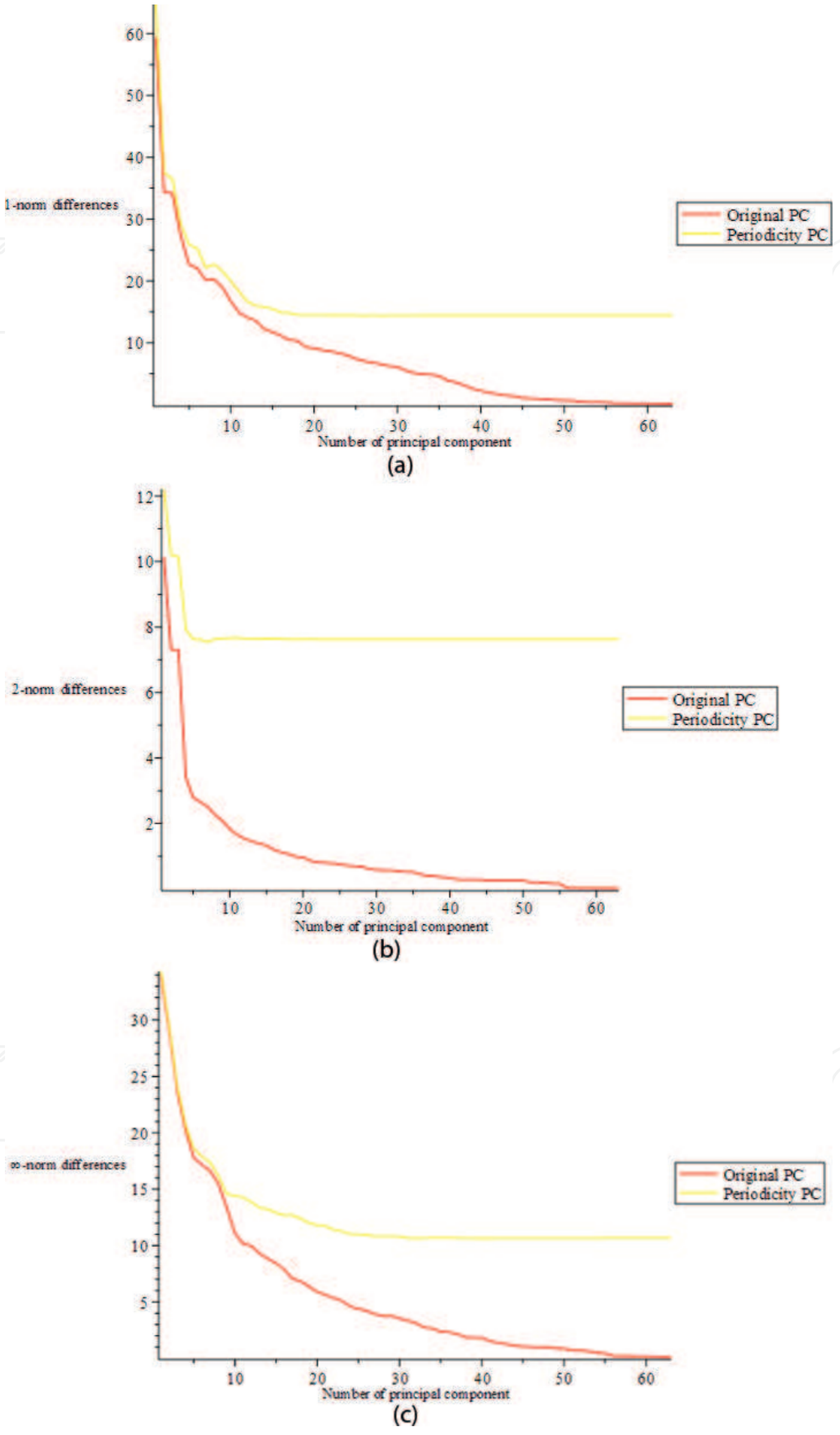

Figure 16. Differences between the image and the reconstruction according to the number of chosen components. (a) 1-norm (b) 2-norm (c) ∞-norm.

5.3. Reduction of the first principal component by periodicity

Using the almost periodicity of the first principal component, we can use less information to obtain acceptable reconstructions of the image. If in the first principal component of dimension 64 we repeat the first eight values periodically and use k principal components to reconstruct the image, we go from a reduction of $k/64$ to another of $[(k-1)+8/64]/64$. **Figure 15** shows both the reconstruction of the image with 2 and 8 original principal components and the reconstruction of the image with 2 and 8 principal components, but with the first component replaced by a vector whose coordinates have period 8, we call this $components_{per}$.

The first $components_{per}$ component is not the true one. Therefore, reconstructions from this set cannot be made with total precision. If we use to compare the 1-norm, 2-norm, and ∞-norm of the image and the corresponding reconstruction, with the original principal components and the principal components using their periodicity, we obtain, by varying the number of used principal components, the results shown in **Figure 16**.

With the original principal components (blue), the original image can be completely reconstructed, while if we use only a few components, in this case 10 or less, approximations similar to the original ones are obtained with $components_{per}$ (green).

6. Conclusions

This chapter has been devoted to give a short but comprehensive introduction to the basics of the statistical technique known as principal component analysis, aimed at its application to image compression. The first part of the chapter was focused on preliminaries, mean vector, covariance matrix, eigenvectors, eigenvalues, and distances. That part finished bringing up the problems that the Euclidean distance presents and highlights the importance of using a statistical distance that takes into account the different variabilities and correlations. To that end, a brief introduction was made to a distance that depends on variances and covariances.

Next, in the second part of the chapter, principal components were introduced and connected with the previously explained concepts. Here, principal components were presented as a particular case of linear combinations of random variables, but with the peculiarity that those linear combinations represent a new coordinate system that is obtained by rotating the original reference system, which has the aforementioned random variables as coordinate axes. The new axes represent the directions with maximum variability and provide a simple description of the structure of the covariance.

Then, the third part of the chapter was devoted to show an application of principal component analysis to image compression. An original image was taken and compressed by using different principal components. The importance of carrying out objective measures of quality reconstructions was highlighted. Also, a novel contribution of this chapter was the introduction to the study of the periodicity of the principal components and to the importance of the reduction of the first principal component by periodicity. In short, a novel construction of principal

components by periodicity of principal components has been included, in order to reduce the computational cost for their calculation, although decreasing the accuracy. It can be said that using the almost periodicity of the first principal component, less information to obtain acceptable reconstructions of the image can be used.

Finally, we would not like to finish this chapter without saying that few pages cannot gather the wide range of applications that this statistical technique has found in solving real-life problems. There is a countless number of applications of principal component analysis to solve problems that both scientists and engineers have to face in real-life situations. However, in order to be practical, it was decided to choose and develop step by step an application example that could be of interest for a wide range of readers. Accordingly, we thought that such an example could be one related to data compression, because with the advancements of information and communication technologies both scientists and engineers need to either store or transmit more information at lower costs, faster, and at greater distances with higher quality. In this sense, one example is image compression by using statistical techniques, and this is the reason why, in this chapter, it was decided to take advantage of statistical properties of an image to present a practical application of principal component analysis to image compression.

Acknowledgements

This work was supported by the Universidad de Las Americas, Ecuador, and the Universidad Politecnica de Madrid, Spain.

Author details

Wilmar Hernandez[1]* and Alfredo Mendez[2]

*Address all correspondence to: wilmar.hernandez@udla.edu.ec

1 Universidad de Las Americas, Quito, Ecuador

2 Universidad Politecnica de Madrid, Madrid, Spain

References

[1] Jackson JE. A User's Guide to Principal Components. John Wiley & Sons; 1991

[2] Diamantaras KI, Kung SY. Principal Component Neural Networks: Theory and Applications. John Wiley & Sons; 1996

[3] Elsner JB, Tsonis AA. Singular Spectrum Analysis: A New Tool in Time Series Analysis. Plenum Press; 1996

[4] Rencher AC. Multivariate Statistical Inference and Applications. 1st ed. Wiley-Interscience; 1997

[5] Flury B. A First Course in Multivariate Statistics. Springer-Verlag; 1997

[6] Gnanadesikan R. Methods for Statistical Data Analysis of Multivariate Observations. 2nd ed. Wiley-Interscience; 1997

[7] Jolliffe IT. Principal Component Analysis. 2nd ed. Springer; 2002

[8] Wichern DW, Johnson RA. Applied Multivariate Statistical Analysis. 6th ed. Pearson Education Limited; 2014

[9] Strang G. Linear Algebra and its Applications. 4th ed. Thomson; 2006

[10] Larson R, Falvo D. Elementary Linear Algebra. 6th ed. Houghton Mifflin Harcourt Publishing Company; 2009

[11] Lay DC, Lay SR, McDonald JJ. Linear Algebra and its Applications. 5th ed. Pearson; 2015