

Appendix A

Principal Component Analysis

This appendix contains a detailed description of Principal Component Analysis (PCA).

A.1 PCA by Maximizing Variance

First we will derive PCA by maximizing the variance in the direction of principal vectors. Let us suppose that we have N M -dimensional vectors \mathbf{x}_j aligned in the data matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$. Let \mathbf{u} be a direction (a vector of length 1) in \mathbb{R}^M . The projection of the j -th vector \mathbf{x}_j onto the vector \mathbf{u} can be calculated in the following way:

$$a_j = \langle \mathbf{x}_j, \mathbf{u} \rangle = \mathbf{u}^T \mathbf{x}_j = \sum_{i=1}^M u_i x_{ij} \quad (\text{A.1})$$

We want to find a direction \mathbf{u} that maximizes the variance of the projections of all input vectors \mathbf{x}_j , $j = 1, \dots, N$.

It follows that the mean of the projections is

$$\bar{a} = \frac{1}{N} \sum_{j=1}^N a_j = \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^M u_i x_{ij} = \sum_{i=1}^M u_i \mu_i \quad (\text{A.2})$$

and the variance is¹:

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{j=1}^N (a_j - \bar{a})^2 = \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^M u_i x_{ij} - \sum_{i=1}^M u_i \mu_i \right)^2 = \\ &= \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^M u_i \hat{x}_{ij} \right)^2 = \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^M \sum_{l=1}^M u_i \hat{x}_{ij} u_l \hat{x}_{lj} = \\ &= \sum_{i=1}^M \sum_{l=1}^M u_i u_l \frac{1}{N} \langle \hat{\mathbf{x}}_i, \hat{\mathbf{x}}_l \rangle = \sum_{i=1}^M \sum_{j=1}^M u_i c_{il} u_l = \mathbf{u}^T \mathbf{C} \mathbf{u} \quad (\text{A.3}) \end{aligned}$$

¹Subscript \mathbf{x}_i denotes i -th *column* vector in the matrix \mathbf{X} , while \mathbf{x}_i denotes i -th *row* vector in the matrix \mathbf{X} .

Here, μ_i is the mean of the i -th row in the data matrix \mathbf{X} and \hat{x}_{ij} is the value of x_{ij} with subtracted μ_i . If the vector $\boldsymbol{\mu}$ contains all row means, thus

$$\boldsymbol{\mu} = [\mu_1, \dots, \mu_M]^T = \frac{1}{N} \sum_{j=1}^N x_j \quad (\text{A.4})$$

then²

$$\hat{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu} \mathbf{1}_{1 \times N} \quad (\text{A.5})$$

and \mathbf{C} is the covariance matrix of \mathbf{X} , thus

$$\mathbf{C} = \frac{1}{N} \hat{\mathbf{X}} \hat{\mathbf{X}}^T \quad (\text{A.6})$$

Our goal is to maximize σ^2 under the constraint that $\|\mathbf{u}\| = 1$. Therefore, by using the technique of Lagrange multipliers, we have to maximize the function

$$F(\mathbf{u}; \lambda) = \mathbf{u}^T \mathbf{C} \mathbf{u} - \lambda (\mathbf{u}^T \mathbf{u} - 1) = \sum_{i=1}^M \sum_{j=1}^M u_i c_{ij} u_j - \lambda \left(\sum_{i=1}^M u_i^2 - 1 \right) \quad (\text{A.7})$$

A closed form solution of this maximization problem can be obtained in the following way:

$$\begin{aligned} \frac{\partial F}{\partial u_l} &= \sum_{j=1}^M c_{lj} u_j + \sum_{i=1}^M u_i c_{il} - \lambda 2u_l = 0; \quad l = 1 \dots M \\ \sum_{i=1}^M c_{li} u_i &= \lambda u_l; \quad l = 1 \dots M \\ \mathbf{C} \mathbf{u} &= \lambda \mathbf{u} \end{aligned} \quad (\text{A.8})$$

Therefore, to find \mathbf{u} and λ that maximize (A.7) we have to compute the eigenvectors and the eigenvalues of the covariance matrix \mathbf{C} . The largest eigenvalue equals the maximal variance, while the corresponding eigenvector determines the direction with the maximal variance.

By performing *eigenvalue decomposition* (EVD) or *singular value decomposition* (SVD) of the covariance matrix \mathbf{C} we can diagonalize \mathbf{C} :

$$\mathbf{C} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T \quad (\text{A.9})$$

in such a way that the orthonormal matrix $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_N] \in \Re^{M \times N}$ contains the eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_N$ in its columns and the diagonal matrix $\boldsymbol{\Lambda} \in \Re^{N \times N}$ contains the eigenvalues $\lambda_1, \dots, \lambda_N$ on its diagonal. We will assume that the eigenvalues and the corresponding eigenvectors are arranged with respect to the descending order of

² $\mathbf{1}_{M \times N}$ denotes a matrix of the dimension $M \times N$, where every element equals 1.

the eigenvalues, thus $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$. Therefore, the most of the variability of the input random vectors is contained in the first eigenvectors. Hence, the eigenvectors are called *principal vectors* (also *principal axes* or *principal directions*).

This approach to calculation of principal vectors is very clear and widely used. However, if the size of the data vector M is very large, which is often the case in the field of computer vision, the covariance matrix $\mathbf{C} \in \mathfrak{R}^{M \times M}$ (see equation A.6) becomes very large and eigenvalue decomposition of \mathbf{C} becomes unfeasible. If the number of input vectors is smaller than the size of these vectors ($N < M$), PCA can be sped up using the following method proposed by Murakami and Kumar [92].

Instead of the covariance (outer product) matrix $\mathbf{C} \in \mathfrak{R}^{M \times M}$ the inner product matrix $\mathbf{C}' \in \mathfrak{R}^{N \times N}$ (divided by the number of the input vectors) is calculated:

$$\mathbf{C}' = \frac{1}{N} \hat{\mathbf{X}}^T \hat{\mathbf{X}} \quad (\text{A.10})$$

The eigenvalues and the eigenvectors of the covariance matrix \mathbf{C} can then be determined from the eigenvalues λ'_i and eigenvectors \mathbf{u}'_i of the matrix \mathbf{C}' as:

$$\lambda_i = \lambda'_i \quad (\text{A.11})$$

$$\mathbf{u}_i = \frac{\hat{\mathbf{X}} \mathbf{u}'_i}{\sqrt{N \lambda'_i}}, \quad i = 1 \dots N \quad (\text{A.12})$$

Note that \mathbf{C}' is much smaller than \mathbf{C} when $N \ll M$. Thus, the eigendecomposition of the $M \times M$ matrix \mathbf{C} has been reduced to the much more feasible eigendecomposition of the $N \times N$ matrix \mathbf{C}' .

A.2 Properties of PCA

The orthonormal matrix \mathbf{U} containing the principal vectors can serve as a linear transformation matrix for projection from the high-dimensional input space to the low-dimensional feature space and vice versa. The columns of \mathbf{U} are the basis vectors of the new low-dimensional coordinate frame expressed with the high-dimensional coordinates. Thus an input vector can be projected into the principal subspace using the transformation matrix $\mathbf{U}^T : \mathfrak{R}^M \rightarrow \mathfrak{R}^N$.

$$\mathbf{a} = \mathbf{U}^T \hat{\mathbf{x}} \quad (\text{A.13})$$

Thus, the coefficients a_j are computed as the projections of the input image onto each principal vector:

$$a_j = \langle \hat{\mathbf{x}}, \mathbf{u}_j \rangle = \sum_{i=1}^M u_{ij} \hat{x}_i, \quad j = 1 \dots N \quad (\text{A.14})$$

All the input vectors contained in the input matrix $\hat{\mathbf{X}}$ can thus be projected as $\mathbf{A} = \mathbf{U}^T \hat{\mathbf{X}}$. Since \mathbf{A} is an orthonormal transformation of the mean centered $\hat{\mathbf{X}}$, the principal components are also centered around zero:

$$\mu_{\mathbf{A}} = \frac{1}{N} \sum_{j=1}^N \mathbf{a}_j = \frac{1}{N} \sum_{j=1}^N \mathbf{U}^T \hat{\mathbf{x}} = \mathbf{U}^T \frac{1}{N} \sum_{j=1}^N \hat{\mathbf{x}} = \mathbf{0} \quad (\text{A.15})$$

Now, let us calculate the correlation matrix of \mathbf{A} :

$$\begin{aligned} \mathbf{C}_A &= \frac{1}{N} \mathbf{A} \mathbf{A}^T = \frac{1}{N} \mathbf{U}^T \hat{\mathbf{X}} (\mathbf{U}^T \hat{\mathbf{X}})^T = \mathbf{U}^T \frac{1}{N} \hat{\mathbf{X}} \hat{\mathbf{X}}^T \mathbf{U} = \\ &= \mathbf{U}^T \mathbf{C} \mathbf{U} = \mathbf{U}^T \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \mathbf{U} = \mathbf{\Lambda} \end{aligned} \quad (\text{A.16})$$

Here, we replaced \mathbf{C} with its diagonalized form (A.9) and considered the orthonormality of \mathbf{U} (thus $\mathbf{U}^T \mathbf{U} = \mathbf{I}$). Therefore, the covariance matrix of the transformed data is the diagonal matrix $\mathbf{\Lambda}$, which contains the eigenvalues on its diagonal. This fact has two important implications. Firstly, it proves that the transformed vectors are uncorrelated. Thus the redundancy caused by correlation between the input vectors has been removed. Secondly, it shows that the variance in the direction of the i -th principal axis (the variance of the i -th principal components) is equal to the i -th eigenvalue λ_i , thus $\frac{1}{N} \sum_{j=1}^N a_{ij}^2 = \lambda_i$.

An important property of the diagonalization (A.9) is that it preserves the trace of the matrix which is being diagonalized [89]. Since the sum of the diagonal elements of the covariance matrix is the sum of variances of the input vectors, this implies that the total variance of the input data has been preserved and equals the sum of all eigenvalues:

$$\begin{aligned} \text{VAR}(\mathbf{X}) &= \sum_{i=1}^M \frac{1}{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T = \sum_{i=1}^M c_{ii} = \\ &= \sum_{i=1}^N \lambda_i = \sum_{i=1}^N \frac{1}{N} \mathbf{a}_i \mathbf{a}_i^T = \text{VAR}(\mathbf{A}) \end{aligned} \quad (\text{A.17})$$

Now we will explain how \mathbf{U} can serve as a transformation matrix for projection of the coefficient vector back into the input space. This operation is called *reconstruction*. The coefficient vector \mathbf{a} is reconstructed using the transformation matrix $\mathbf{U} : \mathbb{R}^N \rightarrow \mathbb{R}^M$:

$$\hat{\mathbf{y}} = \mathbf{U} \mathbf{a} = \sum_{j=1}^N a_j \mathbf{u}_j \quad (\text{A.18})$$

Since N eigenvectors composing $\mathbf{U} \in \mathbb{R}^{M \times N}$ span the same subspace in \mathbb{R}^M as all N input images composing $\mathbf{X} \in \mathbb{R}^{M \times N}$, each input image from \mathbf{X} can be perfectly reconstructed without any reconstruction error. What is more interesting to us, is how well an input image is reconstructed from a subset of principal components only.

To realize this, we first consider how the variance is distributed across the principal axes. This distribution is called the *eigenspectrum* and it is practically a plot of eigenvalues sorted in decreasing order. A typical eigenspectrum is depicted in figure (A.1). As one can observe, most of the variance is contained across the first few eigenvectors. This can also be measured with *energy*, which is defined as a fraction of the total variance. The energy contained in the first k eigenvectors can thus be calculated as

$$en_k = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^N \lambda_i} \quad (\text{A.19})$$

The energy plot obtained from the eigenvalues depicted in figure (A.1.a) is shown in figure (A.1.b). Again, it is evident that most of the energy is contained in a first few eigenvectors already.

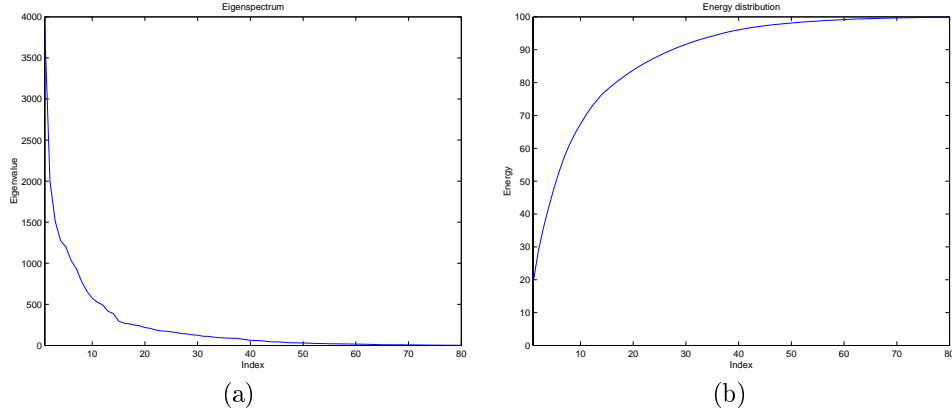


Figure A.1: Typical (a) eigenspectrum, (b) energy distribution.

From this we can conclude that we obtain a good approximation of the input images by considering only a subset of eigenvectors associated with the largest eigenvalues. Henceforth, we will consider only k , $k \ll N$, principal axes, thus $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_k] \in \mathfrak{R}^{M \times k}$.

Now, an input vector is projected into the k -dimensional principal subspace using the transformation matrix $\mathbf{U}^T : \mathfrak{R}^M \rightarrow \mathfrak{R}^k$:

$$\begin{aligned} \mathbf{a} &= \mathbf{U}^T \hat{\mathbf{x}} = \mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}) \\ a_j &= \langle \hat{\mathbf{x}}, \mathbf{u}_j \rangle = \sum_{i=1}^M u_{ij} \hat{x}_i = \sum_{i=1}^M u_{ij} (x_i - \mu_i), \quad j = 1 \dots k \end{aligned} \quad (\text{A.20})$$

and reconstructed using the transformation matrix $\mathbf{U} : \mathfrak{R}^k \rightarrow \mathfrak{R}^M$:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{U} \mathbf{a} = \sum_{j=1}^k a_j \mathbf{u}_j \\ \mathbf{y} &= \hat{\mathbf{y}} + \boldsymbol{\mu} \end{aligned} \quad (\text{A.21})$$

Thus, an input image is approximated with a linear combination of the first k principal vectors.

The reconstruction error (residual error) is equal to the difference between the input and the reconstructed vector:

$$\mathbf{e} = \hat{\mathbf{x}} - \hat{\mathbf{y}} = \sum_{j=1}^N a_j \mathbf{u}_j - \sum_{j=1}^k a_j \mathbf{u}_j = \sum_{j=k+1}^N a_j \mathbf{u}_j \quad (\text{A.22})$$

The most commonly used error measure is the squared reconstruction error, which is defined as a square of the length of the residuum. Considering the orthonormality

of the eigenvectors \mathbf{u}_j we obtain:

$$e = \|\mathbf{e}\|^2 = \left\| \sum_{j=k+1}^N a_j \mathbf{u}_j \right\|^2 = \sum_{j=k+1}^N a_j^2 \quad (\text{A.23})$$

Thus, the squared reconstruction error is equal to the sum of squared discarded principal components. Since they are usually not known, the expected error can be approximated with expected values of the variance across the discarded eigenvectors, which are equal to the corresponding eigenvalues:

$$E(e) = \sum_{j=k+1}^N \lambda_j \quad (\text{A.24})$$

The expected error is thus equal to the sum of the discarded eigenvalues. This consideration confirms the fact that by maximizing the variance in the first (non-discarded) eigenvectors, the squared reconstruction error is being simultaneously minimized. These two assertions are indeed two main properties of PCA.

Therefore, for a given dimension of a subspace k , PCA finds such principal vectors $\mathbf{u}_l, l = 1 \dots k$ and coefficient vectors $\mathbf{a}_j \in \mathfrak{R}^k, j = 1 \dots N$ that minimize the total squared reconstruction error

$$e = \sum_{i=1}^M \sum_{j=1}^N \left(\hat{x}_{ij} - \sum_{l=1}^k u_{il} a_{lj} \right)^2 \quad (\text{A.25})$$

Thus, as an alternative to the maximization of the variance, the principal vectors and the principal components can be estimated by minimizing the squared reconstruction error of equation (A.25). This is a nonlinear minimization problem and can be solved using several proposed algorithms for such a task, e.g., gradient descend algorithm [36] or neural networks [39]. Alternatively, the minimization can be performed by iterating the two-step procedure where first the coefficients are estimated and then the principal vectors are computed. Such an algorithm, which was derived from the probabilistic point of view, can be found in [114].

A.3 Gaussian Interpretation of PCA

In computer vision literature one can find an interesting work [90] where explains a direct connection between Principal Component Analysis (PCA) and the Distance From Feature Space (DFFS) which assumes that the principal subspace can be represented with a Gaussian distribution. As said before, the principal components obtained using PCA preserve the major linear correlations in the data and discard the minor ones. As noted in [90], there are two mutually exclusive and complementary subspaces: the principal subspace (or feature space) that contains the principal components and its orthogonal complementary subspace (the one that is spanned with the discarded components).

As shown in expression (A.23) and (A.25), the residual reconstruction error is defined as

$$e^2(\mathbf{x}) = \sum_{j=k+1}^N a_j^2 = \|\hat{\mathbf{x}}\|^2 - \sum_{i=1}^k a_i^2 \quad (\text{A.26})$$

and can be easily computed from the first k principal components and the L_2 norm of the mean-normalized data vector $\hat{\mathbf{x}}$. The component of \mathbf{x} which lies *in* the feature space F is referred to as the "distance-in-features-space" (DIFS) but is generally not a distance-based norm, but can be interpreted in terms of the probability distribution of the eigenvectors in the feature space. The orthogonal subspace (\bar{F}) is also called "distance-from-feature-space" (DFFS) which is a simple euclidean distance and is equivalent to the residual error $e(x)$ in equation (A.26)

An optimal approach for estimating high-dimensional Gaussian densities is to firstly assume that we have robustly estimated the mean $\bar{\mathbf{x}}$ and covariance Σ of the distribution from a given training set. Under this assumption, the likelihood of an input pattern \mathbf{x} is given by

$$P(\mathbf{x}|\Omega) = \frac{\exp[-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^T \Sigma^{-1}(\mathbf{x} - \bar{\mathbf{x}})]}{(2\pi)^{N/2} |\Sigma|^{1/2}} \quad (\text{A.27})$$

Assuming this Gaussian distribution, the only sufficient statistic for characterizing this likelihood is the *Mahalanobis* distance

$$d(\mathbf{x}) = \hat{\mathbf{x}}^T \Sigma^{-1} \hat{\mathbf{x}} \quad (\text{A.28})$$

where $\hat{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$. Taking in mind that Σ is the covariance matrix of our distribution, we can use the eigenvalues and eigenvectors of Σ and rewrite Σ^{-1} in a diagonalized form

$$d(\mathbf{x}) = \hat{\mathbf{x}} \Sigma^{-1} \hat{\mathbf{x}} \quad (\text{A.29})$$

$$= \hat{\mathbf{x}}^T [\Phi \Lambda^{-1} \Phi^T] \hat{\mathbf{x}} \quad (\text{A.30})$$

$$= \mathbf{a}^T \Lambda^{-1} \mathbf{a} \quad (\text{A.31})$$

where $\mathbf{a} = \Phi^T \hat{\mathbf{x}}$ are the new variables. In other words, a are the eigenvectors of the covariance matrix Σ . So that, the *Mahalanobis* distance can also be expressed in terms of the following sum

$$d(\mathbf{x}) = \sum_{i=1}^N \frac{a_i^2}{\lambda_i} \quad (\text{A.32})$$

Usually, when dealing with high-dimensional data vectors, expression (A.32) is computationally infeasible. We therefore seek to estimate $d(\mathbf{x})$ using only k projections. Intuitively, an obvious choice for a lower-dimensional representation is the principal subspace indicated by PCA which captures the major degrees of statistical variability in the data. Therefore, one can divide the summation into two independent parts corresponding to the principal subspace (the first k projections) and its orthogonal complement subspace (the remaining projections):

$$d(\mathbf{x}) = \sum_{i=1}^k \frac{a_i^2}{\lambda_i} + \sum_{i=k+1}^N \frac{a_i^2}{\lambda_i} \quad (\text{A.33})$$

The first summation can be computed by projecting \mathbf{x} onto the k -dimensional principal subspace F but the remaining terms in the second summation, however, can not be computed explicitly in practice because of the high-dimensionality of data. Nevertheless, the *sum* of these terms is available and is in fact the DFFS quantity $e^2(\mathbf{x})$ which can be computed from expression (A.26). Therefore, based on the available terms, we can formulate an estimator for $d(\mathbf{x})$ as follows

$$\begin{aligned}\hat{d}(\mathbf{x}) &= \sum_{i=1}^k \frac{a_i^2}{\lambda_i} + \frac{1}{\rho} \left(\sum_{i=k+1}^N a_i^2 \right) \\ &= \sum_{i=1}^k \frac{a_i^2}{\lambda_i} + \frac{e^2(\mathbf{x})}{\rho}\end{aligned}\tag{A.34}$$

where the term in the brackets is $e^2(\mathbf{x})$, which as we have seen can be computed using the first k principal components. So, one can write the form of the likelihood based on $\hat{d}(\mathbf{x})$ as the product of two marginal and independent Gaussian densities,

$$\begin{aligned}\hat{P}(\mathbf{x}|\Omega) &= \left(\frac{\exp\left(-\frac{1}{2} \sum_{i=1}^k \frac{a_i^2}{\lambda_i}\right)}{(2\pi)^{k/2} \prod_{i=1}^k \lambda_i^{1/2}} \right) \cdot \left(\frac{\exp\left(-\frac{e^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{(N-k)/2}} \right) \\ &= P_F(\mathbf{x}|\Omega) \hat{P}_{\bar{F}}(\mathbf{x}|\Omega)\end{aligned}\tag{A.35}$$

where $P_F(\mathbf{x}|\Omega)$ is the true marginal density in F -space and $\hat{P}_{\bar{F}}(\mathbf{x}|\Omega)$ is the estimated marginal density in the orthogonal complement \bar{F} -space. Then, the optimal value of ρ can be determined by minimizing a suitable cost function $J(\rho)$. See [90] for more information about this minimizing cost function. The optimal weight ρ^* is determined by [90]

$$\rho^* = \frac{1}{N-k} \sum_{i=k+1}^N \lambda_i\tag{A.36}$$

which is simply the arithmetic average of the eigenvalues in the orthogonal subspace \bar{F} . Then, we can conclude stating that once we select the k -dimensional principal subspace F (as indicated, for example, by PCA), the optimal estimate of the sufficient statistic $\hat{d}(\mathbf{x})$ has the form of expression (A.34) with ρ given by expression (A.36).

One of the typical behaviours in most of the applications it is to simply discard the \bar{F} -space component and simply work with $P_F(\mathbf{x}|\Omega)$. However, the use of the DFFS metric or equivalently the marginal density $P_{\bar{F}}(\mathbf{x}|\Omega)$ is critically important in formulating the likelihood of an observation \mathbf{x} since there are an infinity of vectors which are not members of Ω which can have likely F -space projections.

A.4 Applications of PCA

The most famous application of PCA is the dimensionality reduction in order to formulate the same problem but without noise or with less noise in the data. A

frequently chosen criterion to decide over the dimensionality consists of thresholding the distribution of energy (see figure (A.1.b)) in order to preserve a fixed percentage of the variation. Typical percentages are above 90%. Observing figure (A.1.a) we can see that small variances in the data provided by small eigenvalues are often associated to noise. So, under certain simple assumptions, we can state that PCA reduces noise as well as dimension. Also, notice what happens when eigenvalues exist taking zero values. Dropping their corresponding eigenvectors from the PCA bases has no effect on the mean square error or, what is equivalent, preserves 100% of the data variation, meaning that dimensionality is reduced and no information is lost.

Another application of PCA is data whitening. To whiten or sphere the data is to transform the data linearly so the components of the transformed vector are uncorrelated and have unit variance. The term 'white' comes from the fact that the power spectrum of white noise is constant over all frequencies, resembling the spectrum of white light which contains all colors. Whitening is frequently used as a preprocessing stage, since it can provide invariance to displacement and scale changes. Since PCA uncorrelates the data, one of its applications is to perform whitening. It is also known that statistical independence implies uncorrelation, so PCA can also be understood as a step towards a representation that yields statistically independent components [17]. If the data is Gaussian, the resulting components are in effect independent with unidimensional Gaussian distributions.

Nevertheless PCA is a powerful and simple technique we should not forget the natural limitations derived from its definition, mainly the fact that PCA fails to distinguish high order relationships between the data. This fact should be considered when using PCA as a dimensionality reduction technique previous to classification. Nevertheless, PCA performs successfully in several problems. Two reasons account for this achievement: the first is that in practice it is frequent to find noise in the directions of small variance, the second is that classes are frequently found to be distributed along the main directions of variance.

