# Tutorial on Principal Component Analysis

**Javier R. Movellan**

# 1 Spectral Theorem

Let $A$ be a $k \times k$ positive definite symmetric matrix. Then $A$ can be decomposed as follows:

$$A = P\Lambda P^T \tag{1}$$

where $P$ is a $p \times p$ orthonormal matrix: $PP^T = I$, and $\Lambda$ is a $p \times p$ diagonal matrix.

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda 2 & \cdots & 0 \\ 0 & 0 & 0 & \lambda_p \end{bmatrix} \tag{2}$$

If all $\lambda_i$ are different, $P$ is unique up to multiplication of each column by $-1$ otherwise there is an infinite number of solutions.

## 1.1 Eigen Form

From 1 it follows that

$$AP = P\Lambda \tag{3}$$

or equivalently,

$$Ae_i = \lambda_i e_i \tag{4}$$

where $e_i$ is the $i^{th}$ column of $P$. Thus, *the columns of $P$ are the eigenvectors of $A$*, and the $\lambda_i$ are the eigenvalues associated to each column. For convenience the columns of $P$ are organized such that $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$.

# 2 The Linear Compression Problem

Let

- $X^T = [X_1, \cdots, X_p]$ be a random vector on $(\Omega, \mathcal{F}, P)$, with mean $\mu$ and variance matrix $\Sigma$. Let $\{\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0\}$ be the eigenvalues of the covariance matrix, and $\{e_1, \cdots e_p\}$ the associated eigenvectors.
- $\mathcal{B} = \{u_i \in R^p; i = 1 \cdots k\}$, an orthonormal basis of $\mathbb{R}^k$: $u_i \cdot u_j = \delta_{i,j}$.
- Let $U = [u_1, \cdots, u_k]$ a matrix whose columns are the basis vectors.
- $V$, the subspace spanned by $\mathcal{B}$
- $Y_i = u_i^T X$, the length of the projection of $X$ onto $u_i$. In matrix form: $Y = U^T X$. Call $Y_i$ the $i^{th}$ component of $X$ w.r.t. the basis $\mathcal{B}$
- $\hat{X} = Proj_V X = \sum_{i=1}^k Y_i u_i$, the projection of $X$ onto subspace $V$. In matrix notation $\hat{X} = UY = UU^T X$.

Our goal is to find an orthonormal basis that minimizes the mean square error

$$E\|X - \hat{X}\|^2 = \sum_{i=1}^p E(X_i - \hat{X}_i)^2 \tag{5}$$

### 2.1 Example

$X$ may be a random variable describing a sample of $N$ images. Thus, $x = (x_1, \cdots, x_p)$ represents a specific image. Each component $x_i$ is a pixel value. The distribution of $X$ is defined by the sample: $P(X = x) = 1/N$ if $x$ in is in the sample, 0 otherwise. We want to approximate all the images in the sample as a linear combination of a set of images $\{u_1, \cdots u_k\}$. The basis images have to be orthonormal and we want to choose them s.t. the mean squared error made on a pixel by pixel basis be as small as possible. This gives us a nice compression scheme. The sender and receiver know the basis images. The receiver simply sends the components of a new image (i.e., the inner product between the new image and each of the images in the codebook) and the receiver reconstruct the new image by adding up the images in the codebook times their corresponding weights (i.e., the components).

### 2.2 Network Interpretation

We can frame the problem from a neural net point of view as an auto-encoder problem. An input pattern is transformed into a set of activations in the hidden layer. The output has to be a reconstruction of the input based on the activations of the hidden layer. Let $X^T = [X_1 \cdots X_p]$ is the input to a *linear feed-forward network*. The components $Y = [Y_1 \cdots Y_k]$ are the activations of the hidden layer. The matrix $U = [u_1 \cdots u_p]$ is the matrix of connections between the input and hidden unit. Each vector $u_i$ in the orthonormal basis is the fan-in weight vector from the input $X$ onto the $i^{th}$ hidden unit: $Y_i = u_i^T X = \sum u_{i,j} X_i$. The random vector $\hat{X}$ is the activation of the output layer. The transpose of $U$ is the matrix of connections from hidden units to output units. Thus $\hat{X} = UY = UU^T X$. Our goal is to find a weight matrix $U$ that minimizes the mean squared difference between the input $X$ and the output $\hat{X}$.

The step from input to hidden unit can be seen as an analysis process. The $X$ are modeled as being formed by a combination of *uncorrelated sources*, the components, that we want to recover. The step from hidden to outputs can be seen as a synthesis process. Given the estimated sources, we reconstruct the input.

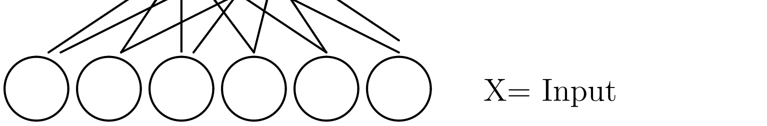

Figure 1: A network interpretation of the compression problem

### 2.3 Fact 1

First of all, using the projection theorem, we know that for a given linear subspace $V$, the best linear approximation to $X$ is the projection of $X$ onto $V$. We symbolize this projection as $\hat{X}$. It follows that,

$$E\|X\|^2 = E\|\hat{X}\|^2 + E\|X - \hat{X}\|^2 \tag{6}$$

**Proof:** For each $x = X(\omega)$,

$$x = \hat{x} + (x - \hat{x}) \tag{7}$$

and since $\hat{x}$ is the projection of $x$ onto subspace $V$,

$$\|x\|^2 = \|\hat{x}\|^2 + \|x - \hat{x}\|^2 \tag{8}$$

Taking expected values, Fact 1 follows.

□

### 2.4 Fact 2

$$E\|\hat{X}\|^2 = \sum_{i=1}^{k} E(Y_i^2) \tag{9}$$

**Proof:**

For each $x = X(\omega)$, the projection of $x$ onto $V$, is

$$\hat{x} = \sum_{i=1}^{k} y_i u_i \tag{10}$$

where $y_i = u_i^T x$ is the length of the projection onto the basis vector $u_i$. Since the $u_i$ are orthogonal to each other,

$$\|\hat{x}\|^2 = \sum_{i=1}^{k} \|y_i u_i\|^2 \tag{11}$$

and since $u_i$ have unit length,

$$\|\hat{x}\|^2 = \sum_{i=1}^{k} y_i^2 \tag{12}$$

Taking expected values, Fact 2 follows.

□

### 2.5 Corollary to Fact 2

Note $E(Y_i) = E(u_i^T X) = u_i^T \mu$. Moreover, $Var(Y_i) = E(Y_i^2) - E^2(Y_i)$. Therefore,

$$E\|\hat{X}\|^2 = \sum_{i=1}^{k} Var(Y_i) - (u_i^T \mu)^2 \tag{13}$$

Thus, if $X$ has zero mean, $\mu^T = [0 \cdots 0]$

$$E\|\hat{X}\|^2 = \sum_{i=1}^{k} Var(Y_i) \tag{14}$$

### 2.6 Remark

Assume the input variables have zero mean $\mu^T = [0 \cdots 0]$, which is always easy to achieve. Using Fact 1 and 2,

$$E\|X - \hat{X}\|^2 = E\|X\|^2 - \sum_{i=1}^{k} Var(Y_i) \tag{15}$$

*Since $E\|X\|^2$ is fixed, minimizing the mean square error is equivalent to maximizing the variance of the components.*

### 2.7 Fact 3

The variance of the coefficients is maximized by using the first $k$ eigenvector of $\Sigma$ as the basis set: $u_i = e_i, i = 1 \cdots k$.

**Proof:**

The proof works by induction. First we assume it is true for $k-1$ and prove it true for $k$. Then we show it is true for $k = 1$.

We know

$$Var(Y_k) = Var(u_k^T X) = u_k^T \Sigma u_k = u_k^T P \Lambda P^T u_k = w^T \Lambda w \tag{16}$$

where $w^T = [w_1 \cdots w_k] = u_k^T P$. Note $w^T \Lambda w$ is a quadratic form, thus

$$Var(Y_k) = \sum_{i=1}^{p} w_i^2 \lambda_i \tag{17}$$

Note $\|w\|^2 = 1$, since $u_k$ has unit length and $P$ is an orthonormal matrix. Now let's assume the first $k-1$ basis vectors are the eigenvector of $\Sigma$, it follows that $w_i = 0, i = 1, \cdots k-1$, because $u_k$ has to be orthogonal to the previous basis vectors. Thus,

$$Var(Y_k) = \sum_{i=k}^{p} w_i^2 \lambda_i \tag{18}$$

Since the $\lambda_i$ are in strictly decreasing order, then the variance is maximized by making $w_k = 1$ and $w_j = 0, j \neq k$. And since $w^T = u_k P$, it follows that $u_k$ has to be $e_k$.

Making $k = 1$ and following the same procedures it is easy to show that $u_1 = e_1$, the first eigenvector, completing the proof.

Note that if some eigenvalues are equal, the solution is not unique. Also if one eigenvalue is zero, there is an infinite number of solutions.

□

### 2.8 Decomposition of Variance

Assuming the inputs have zero mean, then $E\|X\|^2 = \sum_{i=1}^{p} Var(X_i)$. Moreover, from the spectral decomposition of the covariance matrix, we know that $\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} \lambda_i$. Moreover,

$$Var(Y_i) = Var(e_i' X) = e_i' \Sigma e_i = e_i' P \Sigma P e_i = \lambda_i \tag{19}$$

Recalling equation 15, it follows that when our basis is the first $k$ eigenvalues of $\Sigma$,

$$E\|X - \hat{X}\|^2 = \sum_{i=1}^{p} \lambda_i - \sum_{i=1}^{k} Var(Y_i) = \sum_{i=k}^{p} \lambda_i \tag{20}$$

In other words, the error of the approximation equals the sum of the eigenvalues associated with eigenvectors not used in our basis set.

### 2.9 Remarks

- Note that the proof works *regardless of the distribution of* $X$. An interesting aspect to this is that as long as we are projecting onto a linear space, optimal solutions can be achieved using only the mean and covariance matrix of the entire distribution.
- The results do not depend on the fact that the eigenvectors are orthogonal. To see why suppose we are considering a vector $\tilde{u}_k$ which is not-necessarily orthogonal to $u_1 \cdots u_{k-1}$. Then we can choose a vector $u_k$ that is orthogonal to $u_1 \cdots u_k$ and that spans the same space as $u_1 \cdots \tilde{u}_k$. The reconstructions made from the two spaces would be indistinguishable.
- The results depend crucially on the fact that the reconstructions are constrained to be a linear combination of the components of $Y$. To illustrate this fact consider the example in Figure 2. There is a 2 dimensional distribution with equal probability mass at 4 points A, B, C and D. The first eigenvector of this distribution is the vertical axis. Note that the projections of points A and B on the vertical axis are indistinguishable. However the projections of the 4 points on the horizontal axis are distinguishable. A non-linear decoder would be able to perfectly reconstruct those points from the projections on the horizontal axis not from the vertical axis.

### 2.10 Definition of Principal Components

The random variable $Y_i$ is called the $i^th$ principal component of the random vector $X$ iff,

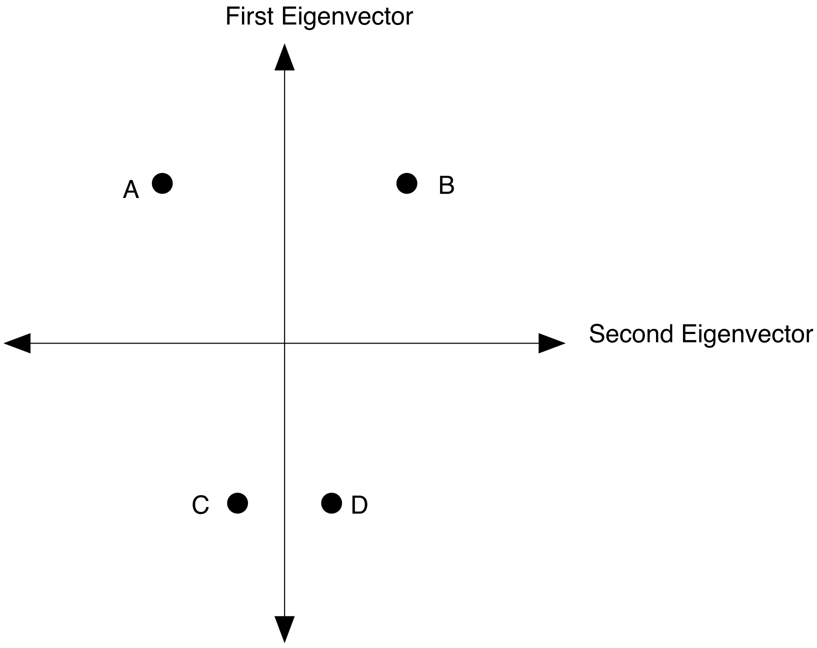

Figure 2: Consider a distribution of 4 equally probably points A, B, C and D. The first principal component in this distribution is the vertical axis. The projections of points A and B on this axis are indistinguishable. However the projections on the horizontal axis are distinguishable. Thus a non-linear decoder would be able to perfectly reconstruct the points using the horizontal projections but not the vertical projections.

1. $Y_k = u'_k X$ where $u_k \in R^p$ and

$$u_k = \underset{u}{\operatorname{argmax}} Var(l'X) \tag{21}$$

2. $Y_i$ is uncorrelated with the previous components $Y_1 \cdots Y_{k-1}$.
3. $u_k$ is orthogonal to the previous basis vectors; $u'_k u_i = \delta_{i,k}, i \leq k$

From the previous sections it follows that if the eigenvalues of $\Sigma$ are different from each other and different from zero, the solution is unique, and $Y_i = e_i X, i = 1 \cdots p$. Otherwise there is an infinite number of solutions. Moreover, the principal components are associated with the mean square error solution only if the mean of $X$ is the zero vector.

### 2.11 Remark

Note that

$$Cov(e_i X, lX) = e'_i \Sigma l = e'_i P \Lambda P' l = \lambda_i e'_i l \tag{22}$$

Assuming the first principal component is $Y_i = e'_i X$, and $\lambda_2 \neq 0$ it follows that making $l$ orthogonal to $e_1$ is equivalent to making $Y_2$ uncorrelated with $Y_1$. However, if $\lambda_2 = 0$, then all solutions, orthogonal and non-orthogonal are uncorrelated.

### 2.12 Interpretation

Principal component analysis models $X$ as a linear combination of uncorrelated hidden sources, which are called the principal components.

If our goal is to decompose $X$ into its underlying hidden sources, we can do so using the following equation:

$$Y_i = e_i X \tag{23}$$

From this point of view, each component of the $i^{th}$ eigenvector tells us how much each of the random variables in $X$ should be weighted to recover the $i^{th}$ "hidden" source.

If our goal is to synthesize $X$ when given a set of underlying sources, we can do so by using the following equation

$$X \approx \hat{X} = \sum_{i=1}^{k} Y_i e_i \tag{24}$$

From this point of view, the $j^{th}$ component of the $i^{th}$ eigenvector tells us the weight of the $i^{th}$ hidden source onto the $j^{th}$ component of X.

### 2.13 Bayesian Interpretation

Section under construction. It will have to do with optimal distribution of X ...

## 3 Implementation issues

Let $X$ be our matrix of data (images in our particular case): $X \in \mathbb{R}^{m \times N}$ where $m$ is the number of images and $N$ is the total number of pixels. In our case then $m \ll N$.
The covariance matrix $C_x$ is defined as

$$C_x = \frac{X'X}{m-1}$$

The matrix of eigenvectors $P$ is such that $C_x = PDP'$ , where $D$ is diagonal and $P$ is orthogonal (since $C_x$ is symmetrical and positive definite).
Now define

$$A = \frac{X'}{\sqrt{m-1}}$$

therefore $C_x = AA'$.

$A$ can be decomposed using singular value decomposition (SVD) in $A = LMO'$ where $L$ and $O$ are orthogonal matrices and $M$ is diagonal. $L \in \mathbb{R}^{N \times N}$, $M \in \mathbb{R}^{N \times m}$, $O \in \mathbb{R}^{m \times m}$. We can then rewrite $C_x$ as:

$$C_x = AA' = LMO'OM'L'$$

since $O$ is orthogonal $\Rightarrow O'O = I \Rightarrow C_x = LMM'L'$.
Comparing this equation with $C_x = PDP' \Rightarrow L \equiv P,\ MM' \equiv D$.

### 3.1 Pentland's shortcut (Turk & Pentland, 1991)

Consider $T = \frac{XX'}{m-1} = A'A$: $T \in \mathbb{R}^{m \times m}$. $T = OM'L'LMO' = OM'MO' \Rightarrow O$ is the matrix of eigenvectors of $T$.

By definition of eigenvectors: if $v_i$ is a generic eigenvector corresponding to an eigenvalue $\lambda_i$:

$$Tv_i = \lambda_i v_i \Rightarrow A'Av_i = \lambda_i v_i \Rightarrow AA'Av_i = A\lambda_i v_i \Rightarrow C_x Av_i = \lambda_i Av_i$$

(the conclusion follows from $C_x = AA'$ and the fact that $\lambda_i$ is a scalar) $\Rightarrow Av_i$ represent the eigenvectors of the original covariance matrix $C_x$. Since $O$ was the matrix of eigenvectors of $T$ then $AO \equiv P$.

In this case we are interested only in the matrix $O$ of the singular value decomposition, then we can use the "economy" version of the SVD in *MATLAB* which computes only the first $m$ columns of $L$ and therefore the first $m$ rows of $M$ giving $L$ of size $N \times m, M$ of size $m \times m$ (the only portion of the original $M$ of size $N \times m$ that was not identically zero) and $O$ exactly as before.

Combining the use of SVD with *Pentland's shortcut* we avoid the multiplication needed to obtain $C_x$ and the computation of all the columns of $L$ and $M$ we are not interested in.

## 4 History

- The first version of this document was written by Javier R. Movellan in 1997 and used in one of the courses he taught at the Cognitive Science Department at UCSD.
- The document was made open source under the GNU Free Documentation License Version 1.2 on October 9 2003, as part of the Kolmogorov project.