A Compact Guide to PCA (CDT-47)

Preprint · December 2020

DOI: 10.13140/RG.2.2.29120.76801/5

1 author:

Luciano da F. Costa
University of São Paulo
734 PUBLICATIONS **13,303** CITATIONS

SEE PROFILE

A Compact Guide to PCA
(CDT-47)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

5th Dec. 2020

Abstract

The statistical linear transform known as Karhunen-Loève transform, as well as the respectively derived principal component analysis methodology, are briefly introduced and illustrated in terms of extremely simple numeric examples. Some important issues regarding the application of PCA, such as standardization, are also addressed.

"Cirrus clouds in the sky, hasty sails on the sea."

LdFC.

1 Introduction

Data to be analyzed, modeled or classified typically involve N observations, individuals or samples, each with M measurements, properties, features of characteristics. Sometimes, a large number of measurements are involved, which not only makes it impossible to visualize the observations, but also impose heavy conceptual and computational limitations.

As a consequence of the tendency of measurements to be correlated, it is possible to reduce the dimension of data as characterized above by using decorrelation techniques allowing redundancy removal. The methodology known as principal component analysis (PCA) provides a simple and effective means to be used for that finality (e.g. [1, 2]).

In this brief work we present the main concepts related to Karhunen-Loève statistical transformation and the principal component analysis methodology in an introductory manner while being focused on practical applications. Two examples are provided regarding the obtention of PCA of data without and with categories.

2 Data Representation

The data to be processed consists of N samples, each being characterized by M measurements, as illustrated in Table 1.

Table 1: A typical representation of the original data to be analyzed: the N rows correspond to samples, and the M columns to respective measurements.

Sample	X_1	X_2	...	X_M
1	2.389	1001.3	...	0.0023
2	-4.764	818.0	...	0.0011
...	...	...	...	...
N	8.490	230.7	...	0.0097

Each of the rows corresponds to a sample, and each of the columns to one of the respective measurements. It is also possible to have the rows to correspond to measurements and the columns to samples. Observe also that each of the measurements may have a specific scale of variation. For instance, measurement X_2 in the example has larger values than X_1, and both these are larger than X_M.

3 Covariance Estimation

The first step in the Karhunen-Loève and PCA application consists in estimating the *covariance matrix* of the data to be analyzed, which can be defined as (e.g. [3]):

$$K_{i,j} = cov(X_i, X_j) \quad (1)$$

where the pairwise covariance being typically estimated

from the data as:

$$cov(X_i, X_j) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{i,k} - \mu_{X_i})(X_{j,k} - \mu_{X_j}) \quad (2)$$

where k indexes each of the N samples.

and μ_{X_i} is the average of measurement X_i and can be estimated as:

$$\mu_{X_i} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (3)$$

Observe that, as a consequence of its definition above, the covariance matrix is always symmetric.

Recall that the covariance between two random variables quantifies the tendency of them to vary together. In case they both tend to increase or decrease jointly, we say that we have a positive correlation. However, if one increases while the other decreases, we have negative correlation. In case there is no joint trend, we say the two variables are uncorrelated (e.g. [3]).

4 Standardization

Given a random variable associated to a measurements X_i, it can be standardized into the respective new random variable $\tilde{X}_i$ as follows:

$$\hat{X}_i = \frac{X_i - \mu_{X_i}}{\sigma_{X_i}} \quad (4)$$

The new variable becomes dimensionless and has null mean and unit variance. In addition, most of the observations of $\tilde{X}_i$ tend to result comprised within the interval $[-2, 2]$.

The choice to standardize or not the random variables associated with a certain data of interest before PCA depends on each specific case. In particular, standardization is typically employed when one or more random variables possesses different ranges of values, so as to avoid larger measurements to have predominating influence on the dispersion (e.g. [4]).

Observe that the covariance matrix estimated over standardized random variables corresponds to the respective matrix of Pearson correlation coefficients (e.g. [3]).

5 The Karhunen-Loève Transform

The Karhunen-Loève transform of the N samples with M measurements each can not be introduced. It involves obtaining the eigenvalues λ_i and respective eigenvectors $\vec{v}_i$, $i = 1, 2, \ldots, M$ of the associated covariance matrix:

$$\begin{array}{cccc} \lambda_1 \geq & \lambda_2 \geq & \ldots \geq & \lambda_M \\ \updownarrow & \updownarrow & \ldots & \updownarrow \\ \vec{v}_1 & \vec{v}_2 & \ldots & \vec{v}_M \end{array} \quad (5)$$

where the eigenvalues are assumed to be sorted in decreasing order.

Given that K is symmetric, its eigenvalues are real and the eigenvectors associated to non-degenerate eigenvalues are orthogonal, hence the matrix performs a rotation of coordinate axes.

We can now stack the obtained eigenvectors as rows of a matrix L, i.e.:

$$L = \begin{bmatrix} \leftarrow & \vec{v}_1 & \rightarrow \\ \leftarrow & \vec{v}_2 & \rightarrow \\ \vdots & \vdots & \vdots \\ \leftarrow & \vec{v}_M & \rightarrow \end{bmatrix} \quad (6)$$

This matrix can be used to implement the Karhunen-Loève transform in terms of the following matrix product:

$$\tilde{\vec{X}} = L\vec{X} \quad (7)$$

So, the original measurements, represented as the random vector $\vec{X}$ are transformed into corresponding new variables in the resulting random vector $\tilde{\vec{X}}$.

The previous matrix equation can be expanded as:

$$\tilde{\vec{X}} = \begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \\ \ldots \\ \tilde{X}_M \end{bmatrix} = \begin{bmatrix} \leftarrow & \vec{v}_1 & \rightarrow \\ \leftarrow & \vec{v}_2 & \rightarrow \\ \vdots & \vdots & \vdots \\ \leftarrow & \vec{v}_M & \rightarrow \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_M \end{bmatrix} \quad (8)$$

Thus, each of the new random variables $\tilde{X}_i$ corresponds to a linear combination, with weights given by the components of the eigenvector $\vec{v}_i$, of the original random variables $\vec{X}$. Therefore, the weights with larger absolute value provide indication about the original variables that contribute more significantly to the overall data variation, providing subsidies for feature selection.

Though similar to the Fourier transform, the Karhunen-Loéve transform has a respective matrix that adapts to each specific data set, hence its optimality. Given that the eigenvectors associated to the largest absolute eigenvalues can be understood as *prototypes* of the data, the new variables $\tilde{X}_i$ correspond to the respective similarity between each data $\vec{X}$ and the prototypes. Remarkably, the new variables $\tilde{X}_i$ can be verified to become necessarily uncorrelated one another, therefore yielding a diagonal respective covariance matrix.

6 Variance Explanation

Each of the eigenvalues λ_i corresponds to the variance of the new random variable $\tilde{X}_i$. Therefore, each of these random variables will account for the following relative contribution to explaining the overall variance of the data:

$$E_i = \frac{\lambda_i}{\sum_{i=1}^{M}\lambda_i} \quad (9)$$

It follows that the explanation of the data variance allowed by the first q new random variables obtained through the Karhunen-Loève transform can then be calculated as:

$$E_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \tag{10}$$

Given that real-world data are often correlated, a good deal of the variance of the data will often be accounted by just a few of the first new variables $\tilde{X}_i$, which are therefore called *principal components*.

In brief, the Karhunen-Loève transform can be understood as performing a rotation of the coordinate systems so that the variance of the original data becomes to a great extent represented in the first new principal components $\tilde{X}_i$, $i = 1, 2, \ldots, q$. Interestingly, it can be shown that this statistical transformation is optimal regarding the compaction of data variance along the first obtained axes (e.g. [1, 2]).

7 Principal Component Analysis

Given that the data variance has been concentrated in the first q new variables (or principal components), it is possible to reduce the dimensionality of the data by leaving out the last $M - q$ axis/variables. The decision on which number q of new variables to preserve depends on each situation, and can consider a minimum percentage of variance explanation. The so-obtained transformation performing dimensionality reduction ($q < M$) is called principal component analysis – PCA.

For instance, In the case of $q = 2$, we have the following PCA:

$$\tilde{\vec{X}} = \begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \end{bmatrix} = \begin{bmatrix} \leftarrow & \vec{v}_1 & \rightarrow \\ \leftarrow & \vec{v}_2 & \rightarrow \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_M \end{bmatrix} \tag{11}$$

8 Example Without Categories

Consider the data involving $N = 3$ samples, with non-specified categories, and $M = 2$ measurements, as indicated in Table 2. Also shown in this table are the average and variance of each of the original measurements X_1 and X_2.

Observe that X_1 has larger values than X_2.

The covariance matrix estimated for the considered data is:

$$K = \begin{bmatrix} 100 & 11 \\ 11 & 1.373 \end{bmatrix}$$

The two eigenvalues obtained are $\lambda_1 = 101.212$ and $\lambda_1 = 0.161$, with respective eigenvectors $\vec{v}_1 = [-0.9940, -0.1095]$ and $\vec{v}_2 = [0.1095, -0.9940]$, from which we get the Karhunen-Loève transformation matrix:

Table 2: Original data for the first example of PCA application.

Sample	X_1	X_2
1	10	1.1
2	20	1.5
3	30	3.3
mean	20.0	1.97
variance	100.0	1.137

$$L = \begin{bmatrix} -0.9940 & -0.1095 \\ 0.1095 & -0.9940 \end{bmatrix}$$

As expected, we have the eigenvectors are orthogonal, i.e.:

$$LL^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The new variables $\tilde{X}_i$ are presented in Table 3. Again, keep in mind that the eigenvectors cannot be specified up to a multiplicative constant, so these results may be different from those calculated by using other implementations. The important point is that, even so, they will be related through a constant scalar value.

Table 3: Original data for the first example of PCA application.

Sample	X_1	X_2
1	-10.060	0.00176
2	-20.044	0.699
3	-30.181	0.00530
mean	20.095	0.235
variance	101.211	0.161

Some interesting features can be observed in these results. First, we have that the variable X_1 became negative. This is explained by the fact that eigenvectors cannot be specified up to a multiplying constant, so inversion of any of the principal axes can be frequently observed.

Second, we have that the second variable $\tilde{X}_2$ has much smaller dispersion than the former $\tilde{X}_1$. Indeed, the variable explanation accounted for just the first principal component can be determined as being 99.84%.

Third, we have that the new variables have similar absolute values as their original counterparts. This is a consequence of the original values of X_1 being much larger than those of X_2, implying the former variable to dominate in defining the minimum variance orientations.

The third effect observed above provides motivation for repeating the PCA on the standardized version of the original measurements, which are presented in Table 4.

Table 4: The standardized version of the data in Table 2.

Sample	X_1	X_2
1	-1	-0.7395
2	0	-0.3982
3	1	1.1377
mean	0.0	0.0
variance	1.0	1.0

In this case, we obtain the following covariance matrix and respectively associated transformation matrix:

$$K = \begin{bmatrix} 1.0 & 0.939 \\ 0.939 & 1.0 \end{bmatrix} \quad L = \begin{bmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$$

The resulting transformed data is shown in Table 5

Table 5: The PCA result considering standardized versions of the data in Table 2, shown in Table 4.

Sample	X_1	X_2
1	1.23	-0.184
2	0.282	0.282
3	-1.511	-0.0974
mean	0.0	0.0
variance	1.939	0.0613

The first important feature of this new version of the original data is that the magnitudes of $\tilde{X}_1$ and $\tilde{X}_2$ are much more similar now, so that neither of them will have a too strong effect on the projection. Interestingly, a high percentage of data variance is provided also in this case, corresponding to 96.93%. As a consequence of the relatively tiny contribution of the second principal component, it may be left out, therefore implying a dimensionality reduction from 2 to 1 without much loss of variance explanation, which corresponds to the main objective in PCA.

9 Example With Categories

Though PCA is a non-supervised statistical method, it is often applied to data in which the samples have respective categories assigned. A simple example of this situation is shown in Table 6.

Table 6: Example of data with categorized samples or individuals.

Category	*Sample*	X_1	X_2
1	1	10	1.1
1	2	20	1.5
1	3	30	3.3
2	4	5.2	3.2
3	5	0.012	0.037
3	6	0.025	0.033
	mean	10.872	1.528
	variance	143.546	2.116

In this situation, PCA is applied as before while ignoring the categories. Let's obtain the PCA for the standardized version of the data in Table 6, which are shown in Table 7. Observe that the standardization is performed on the whole set of samples, not category-by-category.

Table 7: Standardized version of the data in Table 6.

Category	*Sample*	X_1	X_2
1	1	-0.0729	-0.295
1	2	0.762	-0.019
1	3	1.596	1.218
2	4	-0.473	1.149
3	5	-0.906	-1.025
3	6	-0.905	-1.028
	mean	0.0	0.0
	variance	1.0	1.0

In this case, we obtain the following covariance matrix and respective transformation matrix:

$$K = \begin{bmatrix} 1.0 & 0.653 \\ 0.653 & 1.0 \end{bmatrix} \quad L = \begin{bmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$$

The first principal component $\tilde{X}_1$ resulted with much larger variance than the second, yielding a percentage of variance explanation of 82.67%.

It is interesting to observe that identical transformation matrices have been obtained for this as well as the data in the previous section. This is a consequence of the standardization and the low dimensionality (only 2 measurements) of the data, which lead to covariance matrices with identical spectral properties. More diverse transformation matrices could be expected when analyzing higher dimensional data.

10 Concluding Remarks

Data analysis corresponds to an important task frequently employed in science and technology. In the present work, we aimed at providing a brief introduction to the concepts of Karhunen-Loève transform and principal component analysis – PCA.

Special attention was focused on illustrating PCA in terms of simple numeric examples respectively to data single and multiple categories. Toy data with 2 dimensions was purposely adopted in order to allow the intermediate results to be directly compared. Needless to way, PCA is typically applied to much higher dimensional data with more observations.

The important issue of standardization was also addressed and its effect illustrated. Other related concepts and issues were also addressed in order to help PCA applications.

Acknowledgments.

Luciano da F. Costa thanks Henrique F. Arruda for reproducing some of the results and to CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

References

[1] R. A. Johnson and D.W. Wichern. *Applied multivariate analysis*. Prentice Hall, 2002.

[2] F. Gewers, G. R. Ferreira, H. F. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. da F. Costa. Principal component analysis: A natural approach to data exploration. Researchgate, 2019. `https://www.researchgate.net/publication/324454887_Principal_Component_Analysis_A_Natural_Approach_to_Data_Exploration`. accessed 1-Oct-2020.

[3] L. da F. Costa. Multivariate statistical modeling. `https://www.researchgate.net/publication/340442989_Multivariate_Statistical_Modeling_CDT-26`, 2019. [Online; accessed 10-Apr-2020.].

[4] L. da F. Costa. Features transformation and normalization. Researchgate, 2019. `https://www.researchgate.net/publication/340114268_Features_Transformation_and_Normalization_A_Visual_Approach_CDT-24`. [Online; accessed 10-Apr-2020.].

Costa's Didactic Texts – CDTs

CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also incorporate new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

Each CDT focuses on a limited set of interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being relatively short. Links to related material are provided in order to provide some complementation of the covered subjects.

Observe that CDTs, which come with absolutely no warranty, are non distributable and for non-commercial use only.

Please check for new versions of CDTs, as they can be revised. Also, CDTs can and have been cited, e.g. by including the respective DOI. The complete set of CDTs can be found at: `https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs`.