Mathematics for Data Analysis

– Principal Component Analysis –

Matteo Gorgone

University of Messina, Department MIFT

email: mgorgone@unime.it

What is PCA?

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large datasets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Reducing the number of variables of a dataset naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. The data are projected into the Principal Components!

When apply PCA?

PCA is most commonly used when many of the variables are highly correlated with each other and it is desirable to reduce their number to an independent set.

PCA: applications

PCA is used in exploratory data analysis, making predictive models, measuring components of human intelligence, summarise data on variation in human gene frequencies across different regions, market research and finance, neuroscience.
The idea of PCA is simple: reduce the number of variables of a dataset, while preserving as much of the data's variation as possible.

Principal components in Data Analysis

Principal Components are new variables that are constructed as linear combinations of the initial variables. These combinations are done in such a way that the new variables are uncorrelated and most of the information within the initial variables is compressed into the first components.
Organizing information in principal components, will allow you to reduce dimensionality without losing much information. This is achieved by discarding the components with low information and considering the remaining components as your new variables!

Remark

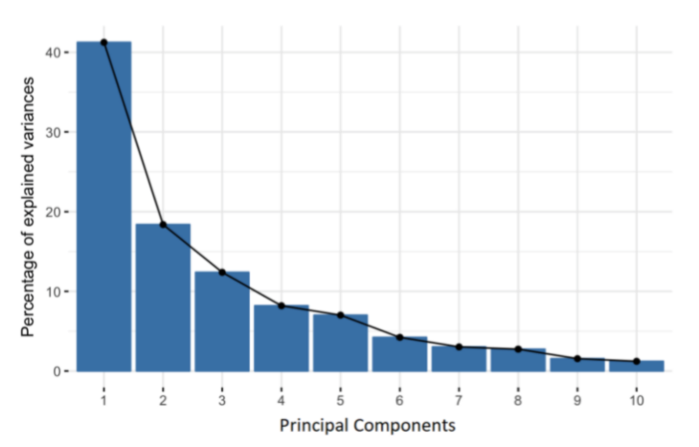
Principal Components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

Intepretation of Principal Components

The first Principal Component of a set of p variables is the derived variable formed as a linear combination of the original variables that explains the largest possible variance in the dataset. The second principal component explains the largest variance in what is left once the effect of the first component is removed, and we may proceed through p iterations until all the variance is explained.

Example

Suppose we have 10-dimensional data (we obtain 10 principal components).
PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like this:



Percentage of variance (information) for each principal component.

Principal Components in Geometry

The Principal Components of a collection of points in a real vector space are a sequence of p unit vectors, where the i-th vector is the direction of a line that best fits the data while being orthogonal to the first $(i-1)$ vectors. In this case, a best-fitting line is defined as one that minimizes the mean squared Euclidean distance from the points to the line.
The directions constitute an orthonormal basis in which different individual dimensions of the data are linearly uncorrelated.
Geometrically speaking, Principal Components represent the directions of the data that explain a maximal amount of variance, that is to say, the lines that capture most information of the data. The relationship between variance and information here, is that, the larger the variance carried by a line, and the larger the dispersion of the data points along it, the more the information it has.

How to compute Principal Components

The first Principal Component can be defined as a direction that maximizes the variance of the projected data. The i-th Principal Component can be computed with the condition that it is uncorrelated with the first $(i-1)$ Principal Components, with direction orthogonal to them, that maximizes the variance of the projected data.
We obtain a number of Principal Components equal to the original number p of variables.

Aim of PCA

Principal Component Analysis is the process of computing the Principal Components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.
Such dimensionality reduction can be a very useful step for visualizing and processing high-dimensional datasets, while still retaining as much of the variance in the dataset as possible.
Essentially, PCA is defined as a linear transformation that transforms the data to a new coordinate system such that the greatest variance of the projected data lies on the first principal component, the second greatest variance on the second principal component, and so on.

PCA: starting point

Suppose we have a data set with N numerical data points $\mathbf{x}_i$ measuring p characters of a population, say

$$\mathbf{x}_i = \{x_{i1}, x_{i2}, \dots, x_{ip}\}, \qquad i = 1, \dots, N.$$

Then, we consider a $N \times p$ matrix containing our data set such that:

- N is the number of the vector data (each row represents a different repetition of the experiment);
- p is the number of features we measured for our phenomenon (thus each measure is a vector $\mathbf{x}_i$ with p components).

PCA: step by step

1. **Standardization.**
The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges, which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.
Mathematically, in order to trasform all the variables to the same scale, we have to subtract the mean and dividing by the standard deviation for each value of each variable:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \qquad i = 1, \dots, N, \qquad j = 1, \dots, p,$$

where x_{ij} is the generic element of the i-th row and j-th column, μ_j and σ_j the mean and the standard deviation of the j-th column (that is the j-feature of the data set).

PCA: step by step

2. **Covariance matrix computation.**
The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix.
The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables $\widetilde{\mathbf{x}}_i$ $(i = 1, \dots, p)$, *i.e.*,

$$\operatorname{cov}(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_j) = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ki} - \mu_i)(x_{kj} - \mu_j), \qquad i, j = 1, \dots, p.$$

Since the covariance of a variable with itself is its variance $(\operatorname{cov}(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_i) = \sigma^2(\widetilde{\mathbf{x}}_i))$, in the main diagonal we actually have the variances of each initial variable.
The sign of a covariance is important:

- if positive then the two variables increase or decrease together (positive correlated);
- if negative then one increases when the other one decreases (negative, or inversely, correlated).

Therefore, the covariance matrix is not more than a table that summaries the correlations between all the possible pairs of variables.

PCA: step by step

3. **Identify Principal Components.**

   Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the Principal Components of the data.

   In fact:

   - the eigenvectors of the Covariance matrix are the directions of the axes where there is the largest variance (most information) and that we call Principal Components;
   - the eigenvalues, corresponding to each eigenvectors, represent the amount of variance carried in each Principal Component.

   By ranking the eigenvectors in order of their associated eigenvalues, highest to lowest, the Principal Components in order of significance are obtained.

PCA: step by step

4. **Feature Vector.**
   Once we computed the eigenvectors and ordered them by their eigenvalues in descending order, we can choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call Feature Vector.
   Then, Feature Vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. This makes it the first step towards dimensionality reduction, because if we choose to keep only q eigenvectors (the Principal Components) out of p, the final dataset will have only q dimensions (q measured features). The Feature Vector has size $p \times q$!

Problem

How to determine an appropriate number of Principal Components?

Procedure based on the average of the eigenvalues

A procedure to determine the number of Principal Components to retain is to choose it for which the eigenvalues are greater than their arithmetic mean.
But, the mean $\overline{\lambda}$ of the eigenvalues of the covariance matrix of the standardized variables is

$$\overline{\lambda} = \frac{1}{p}\sum_{i=1}^{p}\lambda_i = \frac{\mathrm{tr}(\mathrm{Cov})}{p} = 1.$$

In fact, the sum of the eigenvalues turns out to be equal to the trace of the covariance matrix of the standardized variables (the correlation matrix!), *i.e.*, it is equal to p; therefore, the mean eigenvalue is equal to the trace divided by p, that, in this case, is equal to 1.
If a variable has high variance relative to the other variables, the Principal Component will be pulled in the direction of the variable with large variance.

A more efficient Criterion for Principal Components

The amount of the variance that lies along the *i*-th Principal Component can be defined by the explained variance ratio:

$$\text{explained variance ratio of the i-th PC} = \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k}.$$

By summing the explained variance ratio of the first q Principal Components, we obtain the so-called cumulative explained variance:

$$\text{cumulative explained variance of the first q components} = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{k=1}^{p} \lambda_k}$$

Remembering that the eigenvalues are sorted in descending order, a criterion to choose the number q of Principal Components is that the cumulative explained variance is

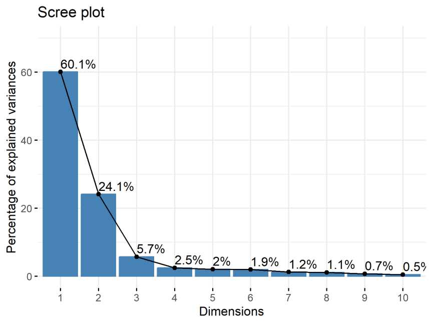
$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{k=1}^{p} \lambda_k} \geq 0.8,$$

i.e., the Principal Components capture at least the 80% of variance of our dataset.

Procedure based on the decrease of the eigenvalues

A graphical procedure is to plot the explained variances (the eigenvalues) $\lambda_1, \lambda_2, \ldots, \lambda_p$ against each Principal Component $1, 2, \ldots, p$. The eigenvalues λ_i are in decreasing order, so one then looks for a dropoff – an elbow – in the curve and retains a number of Principal Components corresponding to the point before the leveling off of the curve, if it does indeed take an elbow shape. Such a plot is called a scree plot, "scree" being the debris at the foot of a glacier (or, more generally, a collection of broken rock fragments at the base of crags, mountain cliffs, volcanoes, or valley shoulders).

Example: decrease of eigenvalues



Based on the scree plot above, the elbow appears at the third component; hence the first three components should be held to derive conclusions.

PCA: step by step

5. **Recast the data along the principal component axes.**
   In the previous steps, apart from standardization, we did not make any changes on the data. We have just selected the Principal Components and formed the Feature Vector, but the input data set are expressed always in terms of the original axes (i.e., in terms of the initial variables).
   In this step, the aim is to use the Feature Vector formed using the eigenvectors of the covariance matrix to reorient the data from the original axes to the ones represented by the Principal Components. This can be done by multiplying the transpose of the original data set by the transpose of the Feature Vector:

$$\mathrm{FinalDataSet}^T = \mathrm{FeatureVector}^T \cdot \mathrm{StandardizedOriginalDataSet}^T,$$

   i.e., we have

$$\mathrm{FinalDataSet} = \mathrm{StandardizedOriginalDataSet} \cdot \mathrm{FeatureVector}.$$

   So doing we compressed the dimension of the data which along the principal axes are represented by a $N \times q$ (with $q < p$) matrix.

Remark

In this step, we are projecting each vector of original data set into the vector space whose basis is the Feature Vector!

Considerations about recasting the data along the principal component axes

In Step 5, we are performing a change of basis and this can have many interpretations:

- FeatureVector^T is the change of basis matrix that transforms StandardizedOriginalDataSet^T into FinalDataSet^T;
- geometrically, FeatureVector^T is a rotation and a stretch which transforms StandardizedOriginalDataSet^T into FinalDataSet^T;
- the rows of FeatureVector^T are a set of new basis vectors (the chosen eigenvectors of the covariance matrix) for expressing the columns of StandardizedOriginalDataSet^T.

Limitations of PCA

- The applicability of PCA is limited to linear correlations between the features but fails when this assumption is violated. Linearity is essential for the change of basis! There are current investigations about nonlinearity, extending the PCA algorithm to that called kernel PCA.
- Both the strength and weakness of PCA is that it is a non-parametric analysis;
- The stardardization process before constructing the covariance matrix could be a limitation. In fact, in fields such as astronomy, all the signals are non-negative, and the stardardization process will force the mean of some astrophysical exposures to be zero, which consequently creates unphysical negative fluxes.
- If PCA is not performed properly, there is a high likelihood of information loss.

Application: Iris flower dataset

The Iris flower dataset is a dataset where you can find collected the sizes of some elements of an Iris flower and a code for three different Iris species. Each flower is described by a 5-dimensional vector: the first component is the sepal length, the second one the sepal width, the third one the petal length, the fourth one the petal width; the last component can be 0 (corresponding to the species Iris setosa), 1 (corresponding to the species Iris versicolor) or 2 (corresponding to the species Iris virginica). We have 150 observations, 50 for each species. Therefore, the dataset is a matrix of order 150×5 (see the external file).

Let us consider only the first 4 columns, and compute their mean values and standard deviations:

$$\begin{aligned} \mu_1 &= 5.84333, & \sigma_1 &= 0.828066,\\ \mu_2 &= 3.054, & \sigma_2 &= 0.433594,\\ \mu_3 &= 3.75867, & \sigma_3 &= 1.76442,\\ \mu_4 &= 1.19867, & \sigma_4 &= 0.763161. \end{aligned}$$

Therefore, the generic element x_{ij} of the i-th row and j-th column is standardized into

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \qquad i = 1, \ldots, 150, \qquad j = 1, \ldots, 4.$$

Application: Iris flower dataset

Now, we can construct the 4 × 4 covariance matrix of the standardized data:

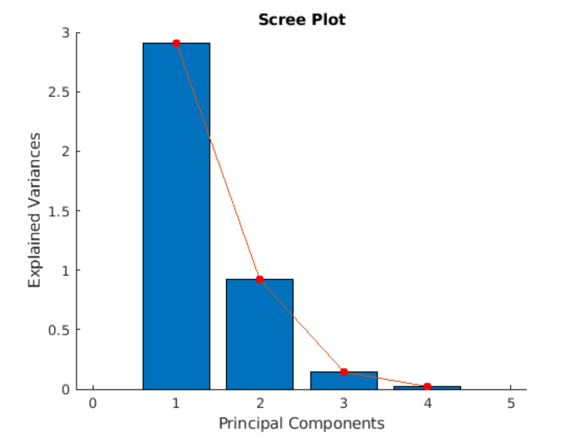
$$C = \begin{bmatrix} 1 & -0.109369 & 0.871754 & 0.817954 \\ -0.109369 & 1 & -0.420516 & -0.356544 \\ 0.871754 & -0.420516 & 1 & 0.962757 \\ 0.817954 & -0.356544 & 0.962757 & 1 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_1 = 2.91082, \qquad \lambda_2 = 0.921221, \qquad \lambda_3 = 0.147353, \qquad \lambda_4 = 0.0206077,$$

sorted in descending order.

Application: Iris flower dataset



This picture is a scree plot that help to interpret the PCA and decide how many components to retain. The start of the bending in the line (point of inflexion) should indicate how many components are retained. In this case, three components should be retained.

Application: Iris flower dataset

The corresponding eigenvectors are

$$\mathbf{p}_1 = \begin{pmatrix} -0.522372 \\ 0.263355 \\ -0.581254 \\ -0.565611 \end{pmatrix}, \qquad \mathbf{p}_2 = \begin{pmatrix} 0.372318 \\ 0.925556 \\ 0.0210948 \\ 0.0654158 \end{pmatrix},$$

$$\mathbf{p}_3 = \begin{pmatrix} -0.721017 \\ 0.242033 \\ 0.140892 \\ 0.633801 \end{pmatrix}, \qquad \mathbf{p}_4 = \begin{pmatrix} 0.261996 \\ -0.124135 \\ -0.801154 \\ 0.523546 \end{pmatrix}.$$

If we define the matrix P whose columns are the four eigenvectors, which diagonalizes the covariance matrix C, then

$$P^T C P = \begin{pmatrix} 2.91082 & 0 & 0 & 0 \\ 0 & 0.921221 & 0 & 0 \\ 0 & 0 & 0.147353 & 0 \\ 0 & 0 & 0 & 0.0206077 \end{pmatrix},$$

i.e., we obtain the diagonal matrix whose entries are the eigenvalues of C.

Application: Iris flower dataset

Since

$$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.95801,$$

we have the the first two principal components (along the directions of vectors $\mathbf{p}_1$ and $\mathbf{p}_2$) capture more than 95% of the total variance of Iris data.

Therefore, defining the 4×2 matrix Q (the Feature Vector) whose columns are the two eigenvectors $\mathbf{p}_1$ and $\mathbf{p}_2$, *i.e.*,

$$Q = \begin{bmatrix} -0.522372 & 0.372318 \\ 0.263355 & 0.925556 \\ -0.581254 & 0.0210948 \\ -0.565611 & 0.0654158 \end{bmatrix},$$

Application: Iris flower dataset

and the 150×4 matrix A of the standardized data (without the last column identifying the species), we can obtain the reduce dataset represented by the 150×2 matrix B such that:

$$B^T = Q^T A^T,$$

i.e.,

$$B = AQ.$$

Denoting by z_{i1}, z_{i2}, z_{i3} and z_{i4} the four elements of the i-th row of the standardized data matrix A, the generic elements b_{ij} of the i-th row of the reduced matrix B are:

$$\begin{aligned} b_{i1} &= -0.522372\, z_{i1} + 0.263355\, z_{i2} - 0.581254\, z_{i3} - 0.565611\, z_{i4}, \\ b_{i2} &= 0.372318\, z_{i1} + 0.925556\, z_{i2} + 0.0210948\, z_{i3} + 0.0654158\, z_{i4}. \end{aligned}$$

Therefore, the reduced standardized data set, according to PCA, is represented by a 150×2 matrix (see the external file).