<https://plot.ly/ipython-notebooks/principal-component-analysis/>

**Intro:**

The sheer size of data in the modern age is not only a challenge for computer hardware but also a main bottleneck for the performance of many machine learning algorithms. The main goal of a PCA analysis is to identify patterns in data; PCA aims to detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense. In a nutshell, this is what PCA is all about: Finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.

**PCA vs LDA**

Both Linear Discriminant Analysis (LDA) and PCA are linear transformation methods. PCA yields the directions (principal components) that maximize the variance of the data, whereas LDA also aims to find the directions that maximize the separation (or discrimination) between different classes, which can be useful in pattern classification problem (PCA "ignores" class labels).  
**In other words, PCA projects the entire dataset onto a different feature (sub) space, and LDA tries to determine a suitable feature (sub) space in order to distinguish between patterns that belong to different classes.**

### PCA and Dimensionality Reduction

Often, the desired goal is to reduce the dimensions of a *d*-dimensional dataset by projecting it onto a (*k*)-dimensional subspace (where *k*<*d*) in order to increase the computational efficiency while retaining most of the information. An important question is "what is the size of *k* that represents the data 'well'?"

Later, we will compute eigenvectors (the principal components) of a dataset and collect them in a projection matrix. Each of those eigenvectors is associated with an eigenvalue which can be interpreted as the "length" or "magnitude" of the corresponding eigenvector. If some eigenvalues have a significantly larger magnitude than others that the reduction of the dataset via PCA onto a smaller dimensional subspace by dropping the "less informative" eigenpairs is reasonable.

### A Summary of the PCA Approach

 Standardize the data.

 obtain the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix, or perform Singular Vector Decomposition.

 Sort eigenvalues in descending order and choose the *k* eigenvectors that correspond to the *k* largest eigenvalues where *k* is the number of dimensions of the new feature subspace (*k* ≤*d*

 )/.

 Construct the projection matrix **W** from the selected *k* eigenvectors.

 Transform the original dataset **X** via **W** to obtain a *k*-dimensional feature subspace **Y**.

## **Iris Dataset**

(<https://archive.ics.uci.edu/ml/datasets/Iris>).

The iris dataset contains measurements for 150 iris flowers from three different species. The three classes in the Iris dataset are:

1. Iris-setosa (n=50)
2. Iris-versicolor (n=50)
3. Iris-virginica (n=50)

And the four features of in Iris dataset are:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

### Exploratory Visualization

In the codes

### Standardizing

Since PCA yields a feature subspace that maximizes the variance along the axes, it makes sense to standardize the data, especially, if it was measured on different scales. Although, all features in the Iris dataset were measured in centimeters, let us continue with the transformation of the data onto unit scale (mean=0 and variance=1), which is a requirement for the optimal performance of many machine learning algorithms.

## **Eigen decomposition Computing Eigenvectors and Eigenvalues**

The eigenvectors and eigenvalues of a covariance (or correlation) matrix represent the "core" of a PCA: The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude. In other words, the eigenvalues explain the variance of the data along the new feature axes.

**Eigen decomposition of the Co-variance or the correlation matrix is done.**

However, the eigendecomposition of the covariance matrix (if the input data was standardized) yields the same results as a eigendecomposition on the correlation matrix, since the correlation matrix can be understood as the normalized covariance matrix.

Eigendecomposition of the standardized data based on the correlation matrix:

### Sorting Eigenpairs

The typical goal of a PCA is to reduce the dimensionality of the original feature space by projecting it onto a smaller subspace, where the eigenvectors will form the axes.

In order to decide which eigenvector(s) can dropped without losing too much information for the construction of lower-dimensional subspace, we need to inspect the corresponding eigenvalues: The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones can be dropped.

In order to do so, the common approach is to rank the eigenvalues from highest to lowest in order choose the top *k* eigenvectors.

### Explained Variance

After sorting the eigenpairs, the next question is "how many principal components are we going to choose for our new feature subspace?" A useful measure is the so-called "explained variance," which can be calculated from the eigenvalues. The explained variance tells us how much information (variance) can be attributed to each of the principal components.

### Projection Matrix

The construction of the projection matrix that will be used to transform the Iris data onto the new feature subspace. Although, the name "projection matrix" has a nice ring to it, it is basically just a matrix of our concatenated top k eigenvectors.

Here, we are reducing the 4-dimensional feature space to a 2-dimensional feature subspace, by choosing the "top 2" eigenvectors with the highest eigenvalues.

### Projection Onto the New Feature Space

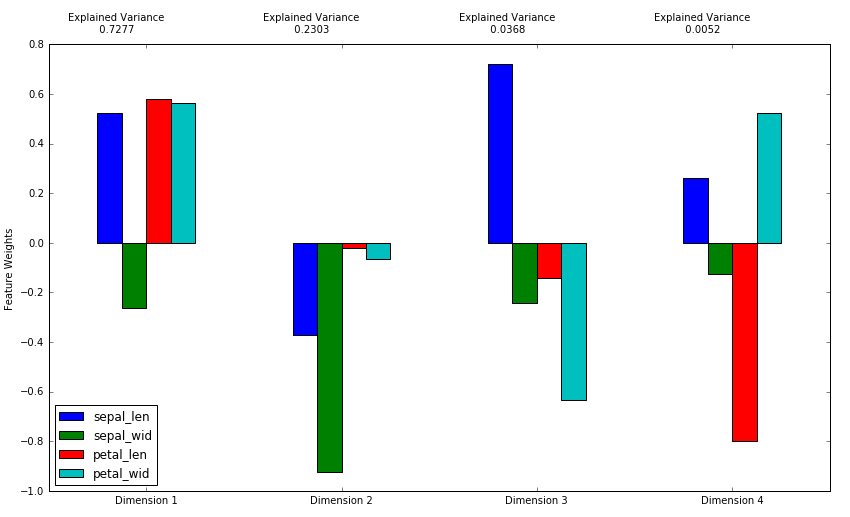
In this last step we will use the 4×2-dimensional projection matrix **W** to transform our samples onto the new subspace via the equation

## **Shortcut - PCA in scikit-learn**

See code

## **Interpretation of Principle Component:-**

To interpret each component, we must compute the correlations between the original data for each variable and each principal component. This is the variance and the principle components of the data.



k1=[] # sepal\_len

k2=[] # sepal\_wid

k3=[] # petal\_len

for i in range(0,len(X\_std1)):

k1.append(X\_std1[i][0])

k2.append(X\_std1[i][1])

k3.append(X\_std1[i][2])

#find correlation

np.corrcoef(k1,reduced\_data['Dimension 1'])

-🡪 0.89122448

np.corrcoef(k3,reduced\_data['Dimension 2'])

🡪0.99

This is as expected from the diagram.

Because of standardization, all principal components will have mean 0. The standard deviation is also given for each of the components and these will be the square root of the eigenvalue.

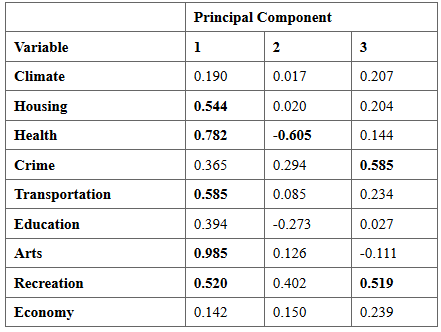
There is zero correlation between the components.

Interpretation of the principal components is based on finding which variables are most strongly correlated with each component, i.e., which of these numbers are large in magnitude, the farthest from zero in either positive or negative direction.

<https://onlinecourses.science.psu.edu/stat505/node/54>

#### **First Principal Component Analysis - PCA1**

The first principal component is strongly correlated with five of the original variables. The first principal component increases with increasing Arts, Health, Transportation, Housing and Recreation scores. This suggests that these five criteria vary together. If one increases, then the remaining ones tend to as well. This component can be viewed as a measure of the quality of Arts, Health, Transportation, and Recreation, and the lack of quality in Housing (recall that high values for Housing are bad). Furthermore, we see that the first principal component correlates most strongly with the Arts. In fact, we could state that based on the correlation of 0.985 that this principal component is primarily a measure of the Arts. It would follow that communities with high values would tend to have a lot of arts available, in terms of theaters, orchestras, etc. Whereas communities with small values would have very few of these types of opportunities.



#### **Second Principal Component Analysis - PCA2**

The second principal component increases with only one of the values, decreasing Health. This component can be viewed as a measure of how unhealthy the location is in terms of available health care including doctors, hospitals, etc.

#### **Third Principal Component Analysis - PCA3**

The third principal component increases with increasing Crime and Recreation. This suggests that places with high crime also tend to have better recreation facilities.

#### **Summary**

* The results of principal component analysis depend on the scales at which the variables are measured.
* Variables with the highest sample variances will tend to be emphasized in the first few principal components.
* Principal component analysis using the covariance function should only be considered if all of the variables have the same units of measurement.