Principal Component Analysis (PCA)

Prerequisite:

- Machine learning fundamentals and feature space
- Linear algebra and statistics

Objectives:

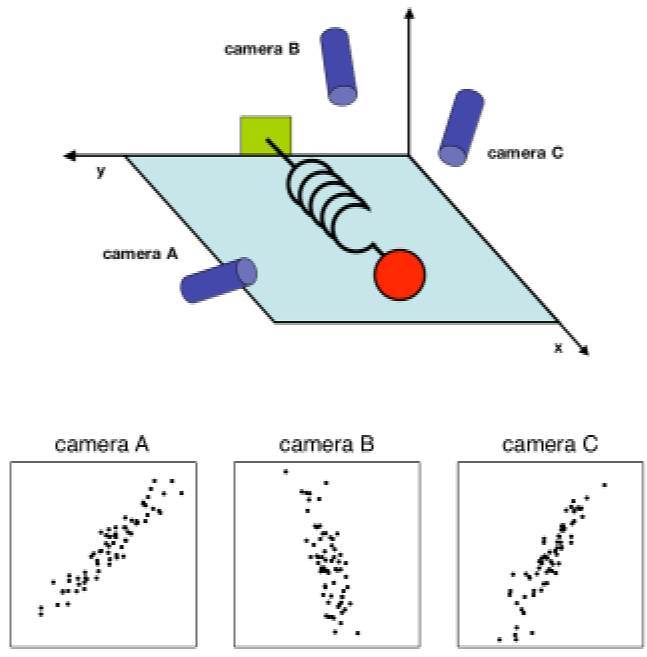
- The motivation for principal component analysis (PCA)
- Understanding what is PCA
- Steps to compute PCA

Motivation and Meaning:

Machine learning model efficiency largely depends on the amount of data. The more and better the data, we will have a better chance to train an efficient predictive model. However, large data also comes with certain limitations like the curse of dimensionality. In high-dimensional data, there is a high possibility to have inconsistency and redundancy in data values. This will pose complexity over computation and analysis of the data. To mitigate this limitation, a technique is proposed to reduce the number of dimensions while selecting the most significant features of the data. This technique is called principal component analysis.

The Principal Component Analysis is a method to identify the pattern in data using similarities and dissimilarities between sample points. The pattern within data is hard to find especially when we cannot visualize it graphically (the data is in high dimension). The principal component analysis is a powerful tool to explore the data with hidden patterns and reduce the number of dimensions. The role of PCA can be understood by a toy example shown in figure 1.

We are interested in studying the motion of the ideal spring. To do that, we have attached a ball of mass m to a frictionless spring and released it at a small distance from the equilibrium. As a result, the spring starts oscillating, and because it is frictionless so it will oscillate indefinitely along the axis at a set frequency. Initially, we don’t know how many dimensions are important, therefore, we record the two-dimensional motion of the ball from three dimensions by placing three cameras around the system and project them. This will generate three different distributions of oscillation and the big question is how do we get a simple equation of x from this data.



The main goal of the Principal Component Analysis (PCA) is to explore a meaningful basis to project the dataset. We are trying to find a new basis that can remove the noise and explore hidden patterns within the data. The intuition to re-express the data to a new basis is understood through figure 2 which shows the data from camera A's point of view. It is clear from the distribution that the largest direction of variation is not along with the original basis of recording (x, y) but along the new line called as best fit line. The figure represents the noise and signals variance by two lines on the data.

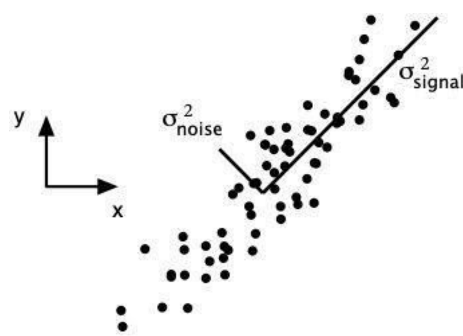


Figure 2 Change of basis (Camera A projection)

So, we can understand principal component analysis (PCA) in many ways like:

- Principal component analysis is a mathematical approach that is used for better interpretation of your data.
- The main purpose of the principal component analysis is to reduce high dimensional data into low dimensional space.

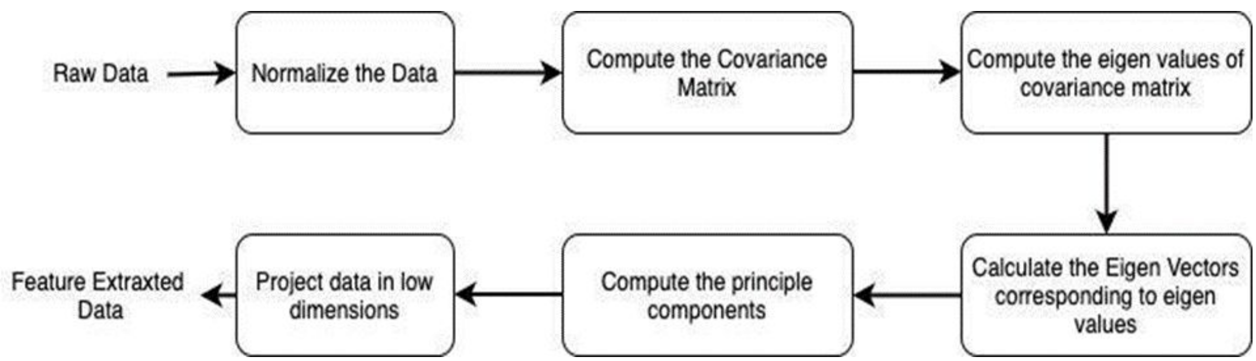
- Principal Component Analysis is an unsupervised machine learning technique which finds insights from the data without having prior knowledge. It reduces the dimension by projecting it geometrically onto a lower-dimensional basis known as principal components.

When to use PCA:

- When you have to reduce the number of features or dimensions in the data.
- When you have to check whether the features are independent of each other or not.

Steps of PCA:

The steps of principal component analysis are summarized as shown in the following chart:



Step 1- Normalization is one of the fundamental steps of data processing. It aims to get an unbiased result from the model. It is the process to bring all the data variables within the same range of values, say (-1, 1). The normal value of the variable is calculated using:

$$Z = \frac{x-\overline{x}}{\sigma}$$

Where $\overline{x}$ is mean and σ is the standard deviation of the distribution.

Step 2- Second step to compute PCA is to calculate the covariance matrix. A covariance matrix describes the correlation between variables within the dataset. It helps to identify all dependent variables. Mathematically a covariance matrix for three features dataset can be defined as:

$$C = [cov(x, x)\ cov(x, y)\ cov(x, z)\ cov(y, x)\ cov(y, y)\ cov(y, z)\ cov(z, x)\ cov(z, y)\ cov(z, z)\ ]$$

Where, $cov\,(x, y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$

- $cov\,(x, y)$ is the variance of variable x w.r.t. y variable

- It is commutative i.e. $cov(x, y) = cov(y, x)$
- $cov(x,x)$ is actually the variance of a variable x

The covariance value within the matrix denotes the dependency of two variables with each other. A negative value denotes that two variables are indirectly proportional to each other and a positive value is for a directly proportional relationship.

Step 3: This step is related to finding the mathematical constructs of the covariance matrix to identify principal components. These mathematical constructs are **eigenvalues** and corresponding **eigenvectors**. An eigenvector is nothing but a principal component of our covariance matrix and represents an axis of new feature space whereas an eigenvalue is the magnitude of that vector. In other words, we can say that eigenvalues actually describe the contribution of each vector in terms of variance (high magnitude for eigenvalue denotes high variance along the corresponding eigenvector in feature space).

Eigenvectors and Eigenvalues- A vector whose direction remains the same even after applying linear transformation is called an eigenvector. This can be expressed as:

$$A.x = \lambda.x$$

Where A is the covariance square matrix, x is an eigenvector and λ is a constant. The eigenvalues of the matrix are obtained by solving the equation-

$$|A - \lambda I| = 0$$

Step 4- Principal components can be computed by arranging eigenvalues with corresponding eigenvectors in descending order. Eigenvectors with large eigenvalues have more significance over the data and form principal components and other eigenvectors can be removed in order to reduce the number of dimensions.

Step 5- Finally the original data can be projected using principal components to reduce the number of dimensions. This can be done by multiplying the transpose of the original dataset with the transpose of computed vectors.

Example- Let's see an example with the help of the steps explained above. We have been given the following data value in two dimensions.

$X1$: 2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2.0, 1.0, 1.5, 1.2

$X2$: 2.4, 0.7, 2.9, 2.2, 3.0, 2.7, 1.6, 1.1, 1.6, 0.9

Step 1- Normalize the value:

$$\overline{X1} = \frac{2.5+0.5+2.2+1.9+3.1+2.3+2.0+1.0+1.5+1.2}{10} = 1.82$$

$$\overline{X2} = \frac{2.4+0.7+2.9+2.2+3.0+2.7+1.6+1.1+1.6+0.9}{10} = 1.91$$

$X1' = X1 - \overline{X1}$ = 0.68, -1.32, 0.38, 0.08, 1.28, 0.48, 0.18, -0.82, -0.32, -0.62

$X2' = X1 - \overline{X2}$ = 0.49, -1.21, 0.99, 0.29, 1.09, 0.79, -0.31, -0.81, -0.31, -1.01

Step 2- Covariance Matrix:

$$C = (0.60 \quad 0.60 \quad 0.60 \quad 0.72\ )$$

Step 3- Calculate the eigenvalue of covariance matrix C.

$$\lambda_1, \lambda_2 = [0.052, 1.26]$$

and the eigenvectors are:

$$V = (-0.74 \ -0.67 \quad 0.67 \ -0.74\ )$$

Step 4- Reduce the dimension and create a new feature vector. To do so, we choose the eigenvector with a bigger eigenvalue (1.26 here) and leave the one with a smaller eigenvalue.

Step 5- Project the data on new feature space-

$$Y = XV^T$$

Y = -0.820, 1.783, -0.988, -0.268, -1.668, -0.907, 0.108, 1.151, 0.445, 1.164

Advantages of PCA

- Removes the correlated attributes
- Help to reduce overfitting
- Improves the data visualization

- It also helps to improve the performance of the algorithm

Disadvantages of PCA

- Data normalization must be needed before applying PCA
- Some level of information loss
- Independent variables become less interpretable