1 Principal Component Analysis

PCA is a method used to reduce number of variables in your data by extracting important one from a large pool. It reduces the dimension of your data with the aim of retaining as much information as possible. In other words, this method combines highly correlated variables together to form a smaller number of an artificial set of variables which is called "principal components" that account for most variance in the data.

As the dimensions of data increases, the difficulty to visualize it and perform computations on it also increases. So, how to reduce the dimensions of a data:-

- Remove the redundant dimensions
- Only keep the most important dimensions

Let us first try to understand some terms:-

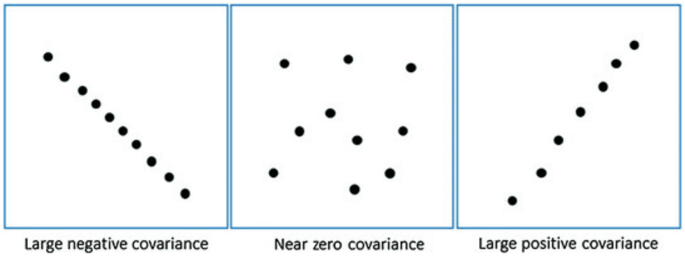
Variance : It is a measure of the variability or it simply measures how spread the data set is. Mathematically, it is the average squared deviation from the mean score. We use the following formula to compute variance var(x).

$$\text{var(x)} = \frac{\sum(x_i - \bar{x})^2}{N}$$

Covariance : It is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction. Formula is shown below denoted by cov(x,y) as the covariance of x and y.

$$\text{var(x)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

- Here, x_i is the value of x in ith dimension.
- $\bar{x}$ and $\bar{y}$ bar denote the corresponding mean values.
- One way to observe the covariance is how interrelated two data sets are.



Positive covariance means X and Y are positively related i.e. as X increases Y also increases. Negative covariance depicts the exact opposite relation. However zero covariance means X and Y are not related.

Now lets think about the requirement of data analysis.

Since we try to find the patterns among the data sets so we want the data to be spread out across each dimension. Also, we want the dimensions to be independent. Such that if data has high covariance when represented in some n number of dimensions then we replace those dimensions with linear combination of those n dimensions. Now that data will only be dependent on linear combination of those related n dimensions. (related = have high covariance)

So, **what does Principal Component Analysis (PCA) do?**

PCA finds a new set of dimensions (or a set of basis of views) such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them. It means more important principle axis occurs first. (more important = more variance/more spread out data)

1.1 How does PCA work?

- Calculate the covariance matrix X of data points.
- Calculate eigenvectors and corresponding eigenvalues.
- Sort the eigenvectors according to their eigenvalues in decreasing order.
- Choose first k eigenvectors and that will be the new k dimensions.
- Transform the original n dimensional data points into k dimensions.

To understand the detail working of PCA, we should have knowledge of eigen values and eigen vectors

Eigenvectors: The directions in which our data are dispersed.

Eigenvalues: The relative importance of these different directions.

[Covariance matrix].[Eigenvector] = [eigenvalue].[Eigenvector]

Lets look into what a covariance matrix is?

A covariance matrix of some data set in 4 dimensions a,b,c,d.

$$\begin{bmatrix} V_a & C_{a,b} & C_{a,c} & C_{a,d} & C_{a,e} \\ C_{a,b} & V_b & C_{b,c} & C_{b,d} & C_{b,e} \\ C_{a,c} & C_{b,c} & V_c & C_{c,d} & C_{c,e} \\ C_{a,d} & C_{b,d} & C_{c,d} & V_d & C_{d,e} \\ C_{a,e} & C_{b,e} & C_{c,e} & C_{d,e} & V_e \end{bmatrix}$$

Va : variance along dimension a

Ca,b : Covariance along dimension a and b

If we have a matrix X of m*n dimension such that it holds n data points of m dimensions, then covariance matrix can be calculated as

$$C_x = \frac{1}{n-1}(X - \bar{X})(X - \bar{X})^T \ X^T = \text{Transpose of X}$$

It is important to note that the covariance matrix contains:-

- variance of dimensions as the main diagonal elements.
- covariance of dimensions as the off diagonal elements.

Also, covariance matrix is symmetric (observe from the image above)

As, we discussed earlier we want the data to be spread out i.e. it should have high variance along dimensions. Also we want to remove correlated dimensions i.e. covariance among the dimensions should be zero (they should be linearly independent).

Therefore, our covariance matrix should have:-

- large numbers as the main diagonal elements.
- zero values as the off diagonal elements.

We call it a diagonal matrix. So, we have to transform the original data points such that their covariance is a diagonal matrix.

Always normalize your data before doing PCA if we use data (features here) of different scales, we get misleading components. We can also simply use correlation matrix instead of using covariance matrix if features are of different scales.

This defines the goal of PCA:-

1. Find linearly independent dimensions which can losslessly represent the data points.
2. Those newly found dimensions should allow us to predict/reconstruct the original dimensions.