**WHAT PCA, WHY PCA, HOW PCA ????????????????**

In real-world data analysis tasks we analyze complex data i.e. multidimensional data. We plot the data and find various patterns in it or use it to train some machine learning models. One way to think about dimensions is that suppose you have a data point x, if we consider this data point as a physical object then dimensions are merely a basis of view, like where is the data located when it is observed from the horizontal axis or vertical axis.

As the dimensions of data increase, 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

First, try to understand some terms -

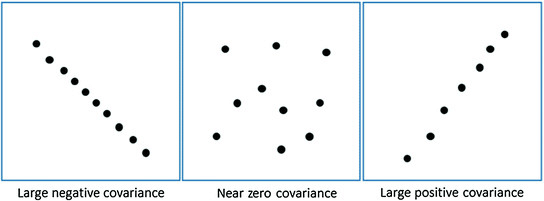
**Variance:** It is a measure of the variability or it simply measures how to spread the data set is. Mathematically, it is the average squared deviation from the mean score. We use the following formula to compute variance *var(x).*

**

**Covariance:** Itis a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction. The formula is shown above denoted by *Cov(x,y)* as the covariance of *x* and *y*.

Here, *xi* is the value of x in *ith* dimension*. x bar* and *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)*

How does PCA work -

1. Calculate the covariance matrix *X* of data points.
2. Calculate eigen vectors and corresponding eigen values.
3. Sort the eigen vectors according to their eigen values in decreasing order.
4. Choose first k eigen vectors and that will be the new k dimensions.
5. Transform the original n dimensional data points into k dimensions.

To understand the detail working of PCA , you should have knowledge of eigen vectors and eigen values. Please refer to this [visual explanation of eigen vectors and values](http://setosa.io/ev/eigenvectors-and-eigenvalues/).