**Things to Remember - Principal Component Analysis ( PCA)**

# **What is PCA?**

Let’s say that you want to predict what the [gross domestic product](http://www.investopedia.com/terms/g/gdp.asp) (GDP) of India will be for 2019. You have lots of information available:. GDP for the first quarter of 2019, GDP for the entirety of 2018, 2017, and so on. You have any publicly-available economic indicators, like the unemployment rate, inflation rate, and so on. You have Census data from 2012 estimating how many Indians work in each industry and Indian statistical data updating those estimates in between each census. You could gather stock price data, the number of [IPOs](https://en.wikipedia.org/wiki/Initial_public_offering) occurring in a year. Despite being an overwhelming number of variables to consider, this just scratches the surface.

With so many variables at hand it would be difficult in hand to decide which variables to focus on, in technical terms it is important to reduce the dimension of your feature space.Reducing the dimension of the feature space is called “dimensionality reduction.

Principal component analysis is a technique for dimension reduction — so it combines input variables in a specific way, to drop the “least important” variables while still retaining the most valuable parts of all of the variables! As an added benefit, each of the “new” variables after PCA are all independent of one another. This is a benefit because the [assumptions of a linear model](http://people.duke.edu/~rnau/testing.htm) require our independent variables to be independent of one another. If we decide to fit a linear regression model with these “new” variables (see “principal component regression” below), this assumption will necessarily be satisfied.

**When should I use PCA?**

1. Do you want to reduce the number of variables, but aren’t able to identify variables to completely remove from consideration?
2. Do you want to ensure your variables are independent of one another?
3. Are you comfortable making your independent variables less interpretable?

If you answered “yes” to all three questions, then PCA is a good method to use. If you answered “no” to question 3, you **should not** use PCA.

**Steps for PCA**

1. Begins by standardising the data. Data on all the dimensions are subtracted from their means to shift the data points to the origin. i.e. the data is centered on the origins
2. Generate the covariance matrix / correlation matrix for all the dimensions
3. Perform eigen decomposition, that is, compute eigenvectors which are the principal components and the corresponding eigenvalues which are the magnitudes of variance captured
4. Sort the eigen pairs in descending order of eigenvalues and select the one with the largest value. This is the first principal component that covers the maximum information from the original data.

**Performance issues with PCA**

1. PCA effectiveness depends upon the scales of the attributes. If attributes have different scales, PCA will pick variable with highest variance rather than picking up attributes based on correlation
2. Changing scales of the variables can change the PCA
3. Interpreting PCA can become challenging due to presence of discrete data
4. Presence of skew in data with long thick tail can impact the effectiveness of the PCA (related to point 1)
5. PCA assumes a linear relationship between attributes. It is ineffective when relationships are non linear