11 February 2018
19:40

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) refers to the process by which principal Component analysis principal components are computed, and the subsequent use of these components in understanding the data. PCA is an unsupervised approach, since it involves only a set of features X1,X2, . . . , Xp, and no associated response Y .

Suppose that we wish to visualize 'n' observations with measurements on a set of 'p' features, ex: X1,X2, . . .,Xp

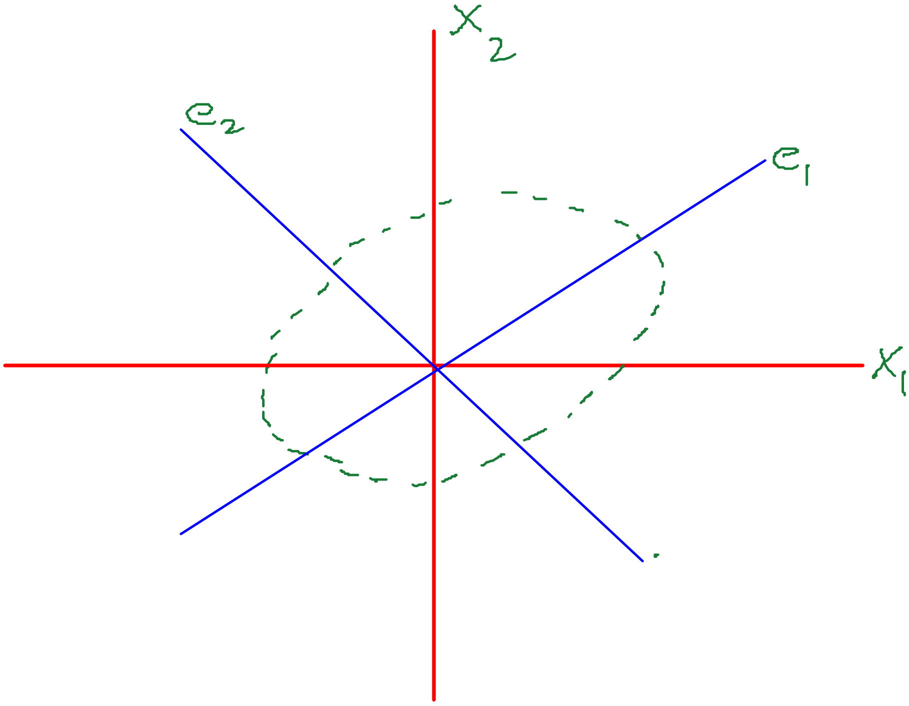
If p is large, then it will certainly not be possible to look at all of them; moreover, most likely none of them will be informative since they each contain just a small fraction of the total information present in the data set. Clearly, a better method is required to visualize the n observations when p is large.

It will be more informative if we have the data in a low dimensional representation rather than p dimensions.

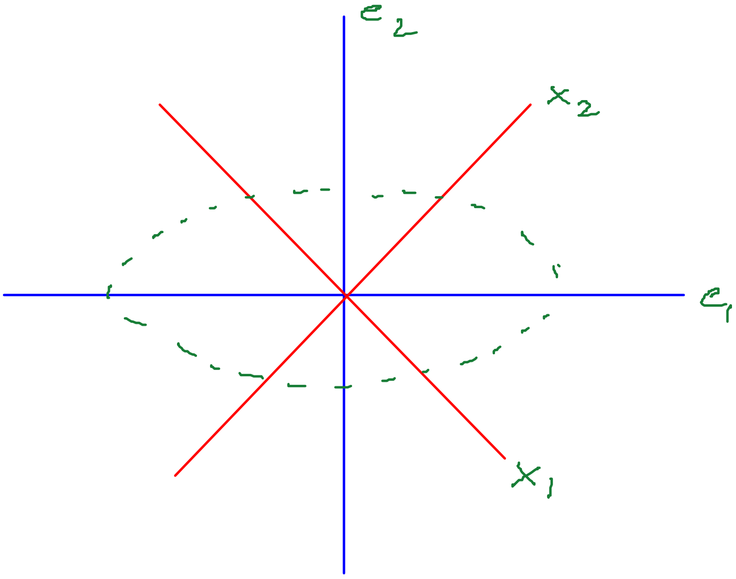
The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting.

PCA helps us to find out the dimensions which are interesting based on the observations how it is varied from each dimensions.

Each of the dimensions found by PCA is a linear combinations of P features.



- In the above, X1 and X2 are correlated and e1 and e2 are the directions
- Make e1 and e2 are the axis. Here e1 and e2 are Eighen vectors.



Now, e1 and e2 are moved to new directions which I am naming as $X1^1$,$X2^1$ as my new variables. These will become independent variables now which don’t have any correlation.

So, When we have a lot of data i.e X1, X2, ...Xn and that will have a lot of correlation among themselves. I want to have variables such that which develops an equation with out having any correlation. So, I am transforming my old variables to new variables which will become independent of each other by using some manipulation.

$X1^1 = X^T e1$
$X2^1 = X^T e2$,....

Initially X1,X2,...Xn are correlated but when converted in to $X1^1$,$X2^1$ it will become independent variables. These new variables are called as Principal Components.

NOTE:

- All the PCA's contains all the original variables since PCA= Linear combination of original variables.
- If we have a dataset contains of multicollinearity issues we can generated some principle components which are independent of each other.
- Can we reduce the new variables of Principle components --> YES
- Principal components are also called as Score data.
- It can be applied when we have more number of continuous and less number of categorical variables.
- X variables need not to be Normal to generate principle components.