Principal component analysis

Principal Component Analysis or PCA is a widely used technique for dimensionality reduction of the large data set. Reducing the number of components or features costs some accuracy and on the other hand, it makes the large data set simpler, easy to explore and visualize. Also, it reduces the computational complexity of the model which makes machine learning algorithms run faster. It is always a question and debatable how much accuracy it is sacrificing to get less complex and reduced dimensions data set. we don’t have a fixed answer for this however we try to keep most of the variance while choosing the final set of components.

**Steps Involved in PCA**

1. Standardize the data. (with mean =0 and variance = 1)
2. Compute the Covariance matrix of dimensions.
3. Obtain the Eigenvectors and Eigenvalues from the covariance matrix (we can also use correlation matrix or even Single value decomposition, however in this post will focus on covariance matrix).
4. Sort eigenvalues in descending order and choose the top k Eigenvectors that correspond to the k largest eigenvalues (k will become the number of dimensions of the new feature subspace k≤d, d is the number of original dimensions).
5. Construct the projection matrix W from the selected k Eigenvectors.
6. Transform the original data set X via W to obtain the new k-dimensional feature subspace Y.

## **What are principal components ?**

A principal component is a normalized linear combination of the original predictors in a data set. In image above, PC1 and PC2 are the principal components. Let’s say we have a set of predictors as X¹, X²...,Xp

The principal component can be written as:

Z¹ = Φ¹¹X¹ + Φ²¹X² + Φ³¹X³ + .... +Φp¹Xp

where,

* Z¹ is first principal component
* Φp¹ is the loading vector comprising of loadings (Φ¹, Φ²..) of first principal component. The loadings are constrained to a sum of square equals to 1. This is because large magnitude of loadings may lead to large variance. It also defines the direction of the principal component (Z¹) along which data varies the most. It results in a line in p dimensional space which is closest to the n observations. Closeness is measured using average squared euclidean distance.
* X¹..Xp are normalized predictors. Normalized predictors have mean equals to zero and standard deviation equals to one.

Therefore,

**First principal component** is a linear combination of original predictor variables which captures the maximum variance in the data set. It determines the direction of highest variability in the data. Larger the variability captured in first component, larger the information captured by component. No other component can have variability higher than first principal component.

The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line.

Similarly, we can compute the second principal component also.

**Second principal component** (Z²) is also a linear combination of original predictors which captures the remaining variance in the data set and is uncorrelated with Z¹. In other words, the correlation between first and second component should is zero. It can be represented as:

Z² = Φ¹²X¹ + Φ²²X² + Φ³²X³ + .... + Φp2Xp

If the two components are uncorrelated, their directions should be orthogonal (image below). This image is based on a simulated data with 2 predictors. Notice the direction of the components, as expected they are orthogonal. This suggests the correlation b/w these components in zero.

# If the two components are uncorrelated, their directions should be orthogonal (image below). This image is based on a simulated data with 2 predictors. Notice the direction of the components, as expected they are orthogonal. This suggests the correlation b/w these components in zero.

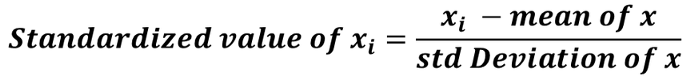
# PCA : Orthogonality of Principal Componentssni

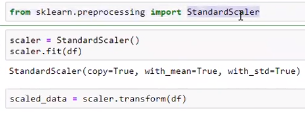
# All succeeding principal component follows a similar concept i.e. they capture the remaining variation without being correlated with the previous component. In general, for n × p dimensional data, min(n-1, p) principal component can be constructed.

# 1. Standardization

When there are different scales used for the measurement of the values of the features, then it is advisable to do the standardization to bring all the feature spaces with mean = 0 and variance = 1.

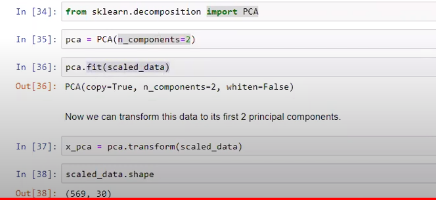
The reason why standardization is very much needed before performing PCA is that PCA is very sensitive to variances. Meaning, if there are large differences between the scales (ranges) of the features, then those with larger scales will dominate over those with the small scales.







N\_components = number of principal components in resultant set



In original dataset we have 3o feature and after PCA it has been reduced to 2

