Principal Component Analysis (PCA)

If data is represented using rows and columns, such as in a spreadsheet, then the input variables are the columns that are fed as input to a model to predict the target variable. Input variables are also called features.

We can consider the columns of data representing dimensions on an *n*-dimensional feature space and the rows of data as points in that space. This is a useful geometric interpretation of a dataset.

Having a large number of dimensions in the feature space can mean that the volume of that space is very large, and in turn, the points that we have in that space (rows of data) often represent a small and non-representative sample.

This can dramatically impact the performance of machine learning algorithms fit on data with many input features, generally referred to as the "curse of dimensionality."

Therefore, it is often desirable to reduce the number of input features. This reduces the number of dimensions of the feature space, hence the name "*dimensionality reduction*."

Principal Component Analysis (PCA) is a **linear dimensionality reduction** technique that can be utilized for extracting information from a high-dimensional space by projecting it into a lower-dimensional sub-space. It tries to preserve the essential parts that have more variation of the data and remove the non-essential parts with fewer variation.

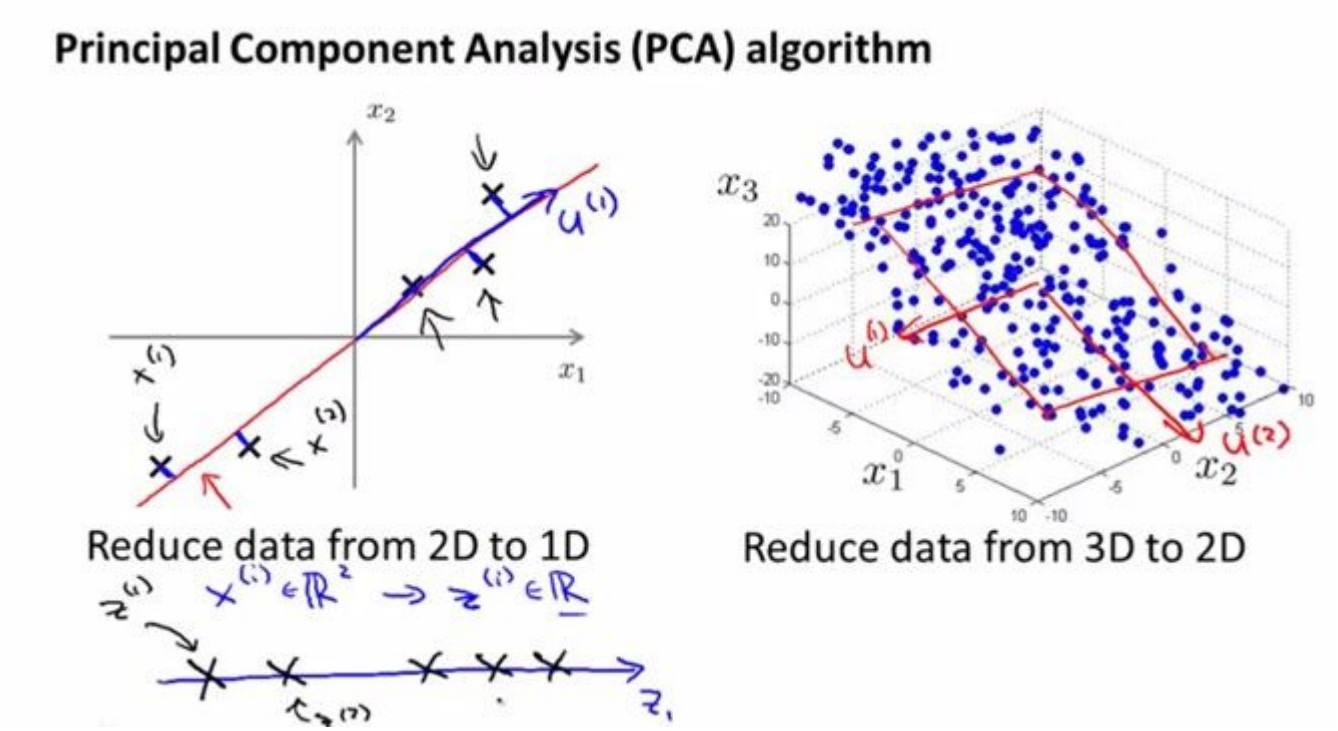

Image Source: Machine Learning Lectures by Prof. Andrew NG at Stanford University

One important thing to note about PCA is that it is an **Unsupervised** dimensionality reduction technique, you can cluster the similar data points based on the feature correlation between them without any supervision (or labels).

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

Note: Features, Dimensions, and Variables are all referring to the same thing. You will find them being used interchangeably.

Features / Attributes / Variables →

Samples ↓

	sepal-length	sepal-width	petal-length	petal-width
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

What is a Principal Component

Principal Component Analyis is basically a statistical procedure to convert a set of observation of possibly correlated variables into a set of values of linearly uncorrelated variables.

In a layman term, when the data is projected into a lower dimension (assume three dimensions) from a higher space, the three dimensions are nothing but the three Principal Components that captures (or holds) most of the variance (information) of data.

Principal components have both direction and magnitude. The direction represents across which *principal axes* the data is mostly spread out or has most variance and the magnitude signifies the amount of variance that Principal Component captures of the data when projected onto that axis. The principal components are a straight line, and the first principal component holds the most variance in the data. Each subsequent principal component is orthogonal to the last and has a lesser variance. In this way, given a set of x correlated variables over y samples you achieve a set of u uncorrelated principal components over the same y samples.

Properties of Principal Component

Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables. The output of PCA are these principal components, the number of which is less than or equal to the number of original variables. Less, in case, when we wish to discard or reduce the dimensions in our dataset. The PCs possess some useful properties which are listed below:

1. The PCs are essentially the linear combinations of the original variables, the weights vector in this combination is actually the eigenvector found which in turn satisfies the principle of least squares.
2. The PCs are orthogonal, as already discussed.
3. The variation present in the PCs decrease as we move from the 1st PC to the last one, hence the importance.

The least important PCs are also sometimes useful in regression, outlier detection, etc.

Uses of PCA

1. It is used to find inter-relation between variables in the data.
2. It is used to interpret and visualize data.
3. As number of variables are decreasing it makes further analysis simpler.
4. It's often used to visualize genetic distance and relatedness between populations.

These are basically performed on square symmetric matrix. It can be a pure sum of squares and cross products matrix or Covariance matrix or Correlation matrix. A correlation matrix is used if the individual variance differs much.

Objectives of PCA

5. It is basically a non-dependent procedure in which it reduces attribute space from a large number of variables to a smaller number of factors.
6. PCA is basically a dimension reduction process but there is no guarantee that the dimension is interpretable.
7. Main task in this PCA is to select a subset of variables from a larger set, based on which original variables have the highest correlation with the principal amount.

Principal Axis Method

PCA basically search a linear combination of variables so that we can extract maximum variance from the variables. Once this process completes it removes it and search for another linear combination which gives an explanation about the maximum proportion of remaining variance which basically leads to orthogonal factors. In this method, we analyze total variance.

Eigenvector:

It is a non-zero vector that stays parallel after matrix multiplication. Let's suppose x is eigen vector of dimension r of matrix M with dimension r*r if Mx and x are parallel. Then we need to solve Mx=Ax where both x and A are unknown to get eigen vector and eigen values. Under Eigen-Vectors we can say that Principal components show both common and unique variance of the variable. Basically, it is variance focused approach seeking to reproduce total variance and correlation with all components. The principal components are basically the linear combinations of the original variables weighted by their contribution to explain the variance in a particular orthogonal dimension.

Eigen Values:

It is basically known as characteristic roots. It basically measures the variance in all variables which is accounted for by that factor. The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables. If the factor is low then it is contributing less in explanation of variables. In simple words, it measures the amount of variance in the total given database accounted by the factor. We can calculate the factor's eigen value as the sum of its squared factor loading for all the variables.

Algorithm

1. *Find the mean vector.*
2. *Assemble all the data samples in a mean adjusted matrix.*
3. *Create the covariance matrix.*
4. *Compute the Eigen vectors and Eigen values.*
5. *Compute the basis vectors.*
6. *Represent each sample as a linear combination of basis vectors.*

Referance

1. Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Publication
2. http://kiwi.bridgeport.edu/cpeg540/PrincipalComponentAnalysis_Tutorial.pdf
3. https://builtin.com/data-science/step-step-explanation-principal-component-analysis
4. https://en.wikipedia.org/wiki/Principal_component_analysis#Computing_PCA_using_the_covariance_method
5. https://www.geeksforgeeks.org/principal-component-analysis-with-python/
6. https://www.datacamp.com/community/tutorials/principal-component-analysis-in-python