PCA (PRINCIPAL COMPENENT 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 subspace. It keeps the data which have higher variance and removes the lower variance data. It helps to find most significant features in dataset. PCA helps to find a sequence of linear combination of variables

The idea of PCA is simple — **reduce the number of variables of a data set, while preserving as much information as possible**.

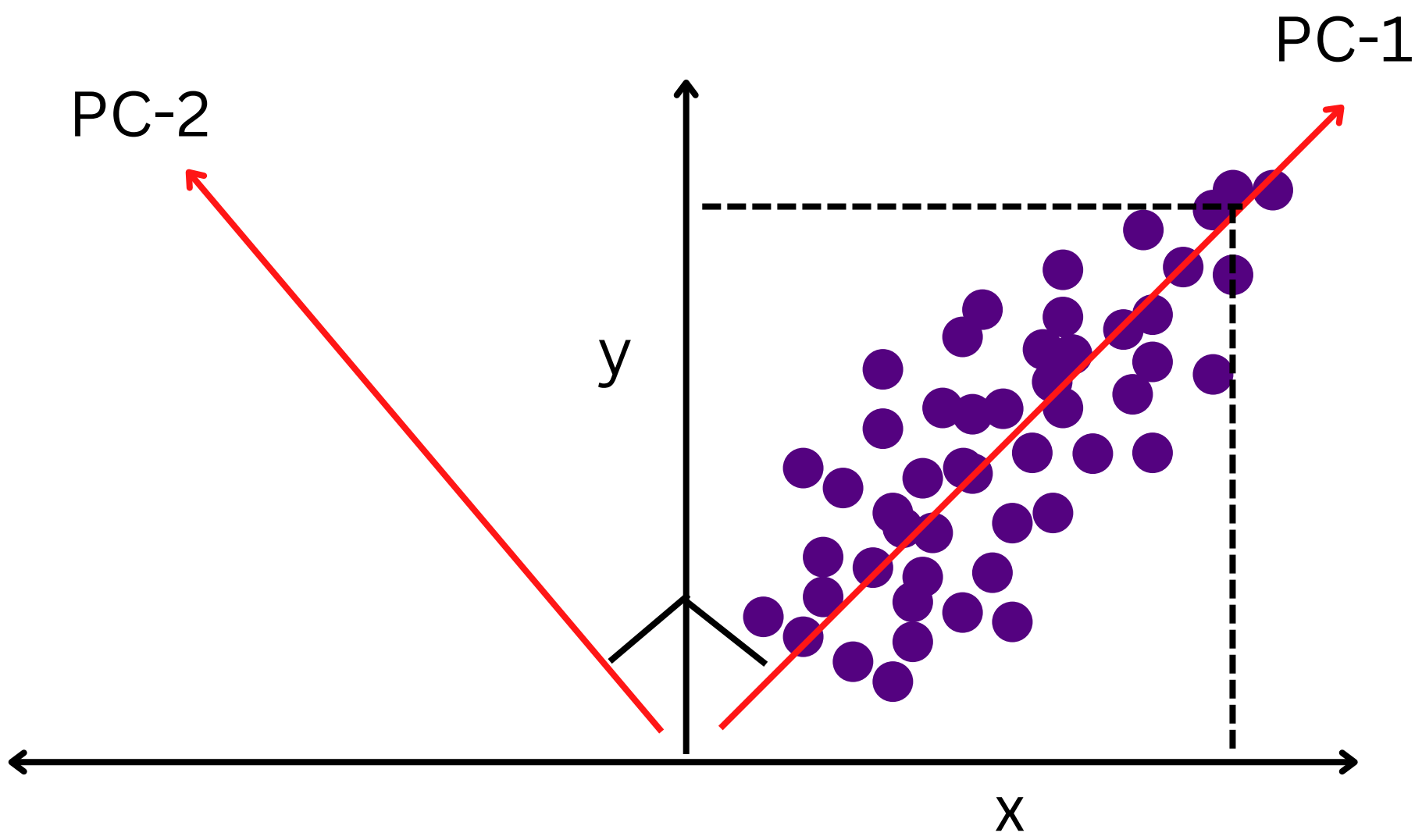

FIG-1.1

We have several points plotted on a 2d plane above fig, There are two Principal Components, **PC-1 is a Primary Principal Component** that explains the maximum variance in the data, **PC-2 is another Principal Component** that is orthogonal to PC-1.

Principal Component: PC are a Straight Line that captures most of the variance of the data. They have "Direction" and "Magnitude". PC are orthogonal projections (Perpendiculars) of the data into lower-dimensional space.

Applications of PCA:

- PCA is used to visualize multidimensio-
-nal data
- It used to reduce the number of dimen-sions in dataset.
- PCA can help to resize an image.
- It also used in Computer Vision, Image compression etc..

How does PCA works

1) Normalize (or) Standardization of the data:

Standardize the data before performing PCA. This will ensure that each feature has mean=0 and variance =1. The ultimate goal of this step is to standardize the range of the continuous initial variables so they each data point contribute equally to the analysis(Feature Scaling).

Why we need standardization prior to PCA. If there are large difference between the ranges of initial variables, the larger ranges will dominate over the smaller ranges, ex: variable ranges between 0 and 100 will dominate over a variable between 0 and 1, Which leads to biased results.

Transforming the data to comparable scales can prevent this problem

$$z = \frac{\textbf{value - mean}}{\textbf{Standard-Deviation}}$$

2) Covariance Matrix Computation: The aim of this step is to understand how the variables of input dataset are varying from the mean with respect to each other. To find is there any relationship between them. Sometimes variables are highly correlated in such a way that they contain redundant information. To identify these correlations, we compute the Covariance Matrics.

3) Eigenvectors and Eigenvalues: These are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the Principal Components of the data. These always comes in pairs, So every “Eigenvectors” has a “Eigenvalues”. And this number is equal to the number of dimensions of data.

4) Feature Vector: Computing the eigenvectors and ordering them by their eigenvalues in descending order, allows us to find the Principal Components or discard those which have lesser significances, and from with the remaining ones a matrix of vectors that we call Feature Vector.

5) Recast the data along with the principal components axes: From Standardization, You do not make any changes on the data you just select the principal component and form the feature vector, but input dataset always remains in terms of original axes.

T= Transpose

$$\textbf{FinalDataset} = \textbf{FeatureVector}^{T} * \textbf{Standardize}\ \textbf{StandardizeOriginalDataset}^{T}$$