PRINCIPAL COMPONENT ANALYSIS

3 DIMENSION REDUCTION

II FEATURE

a Feature selection

Feature selection is the process of **selecting the most relevant features** among your existing features.

To keep "relevant" features only, we will remove features that are:

i. Non informative

ii. Non discriminative

iii. Redundant

3 DIMENSION REDUCTION

II FEATURE

a Feature selection

iii REMOVE **REDUNDANT** FEATURES

Method: High correlation filter

Principle: We remove features that are similar or highly correlated with other feature(s).

Ex: **Same size** in square meters and square inches

Your model doesn't need the **same information twice!**

You can detect **correlated features** computing the Pearson product-moment correlation coefficients matrix.

NumPy

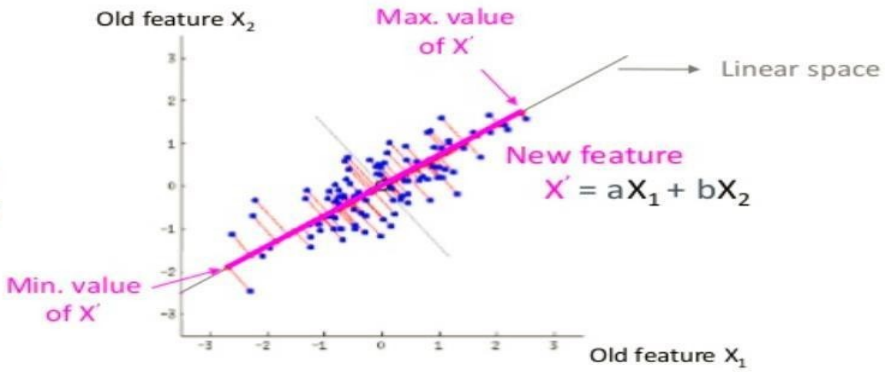
```
import numpy as np
matrix = np.corrcoef(X)
```

b Feature extraction

The most common algorithm for feature extraction is **Principal Component Analysis (PCA).**

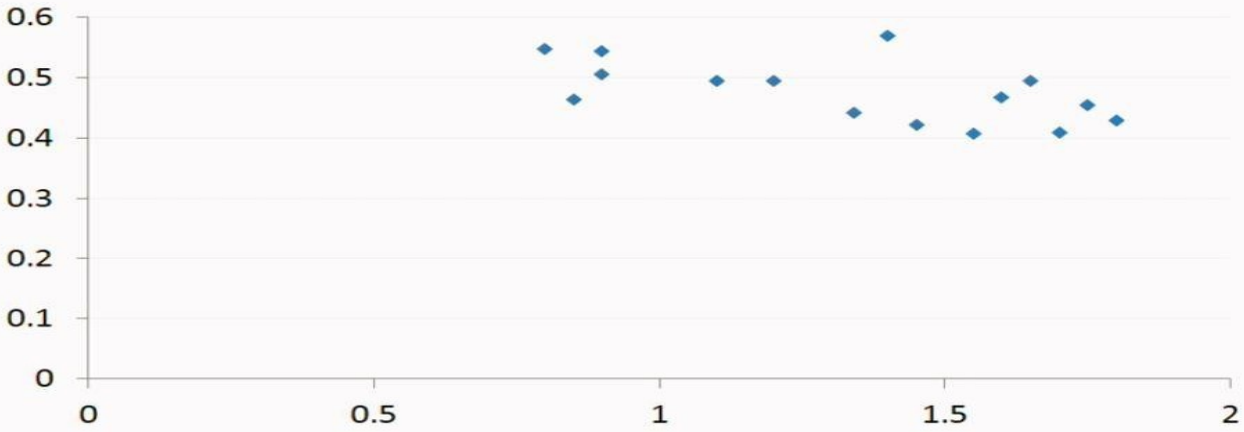
PCA makes an orthogonal **projection on a linear space** to determine new features, called **principal components**, that are a linear combination of the old ones.

Example of reduction of 2 features into a single one



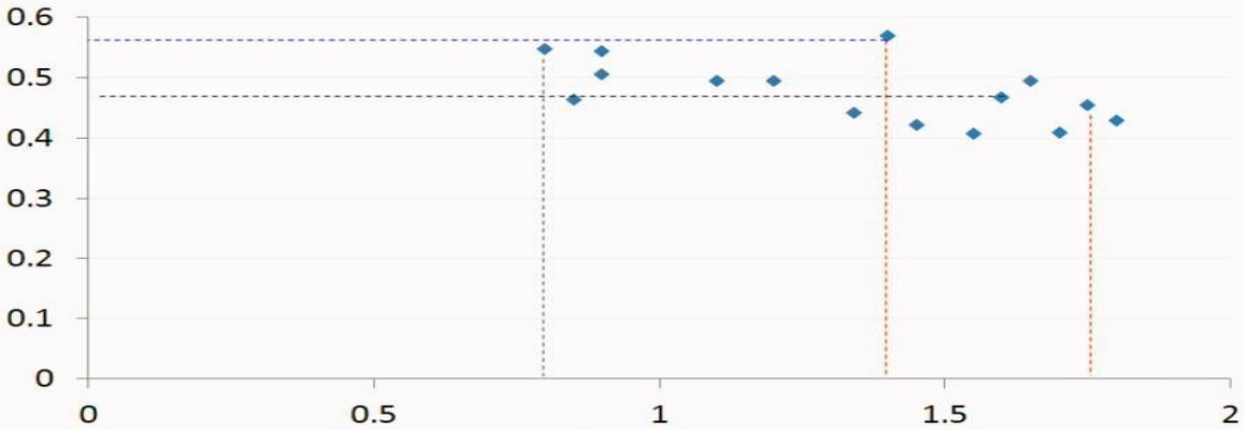
UNDERSTAND THE VARIANCE IN DATA

Which dimension is varying more



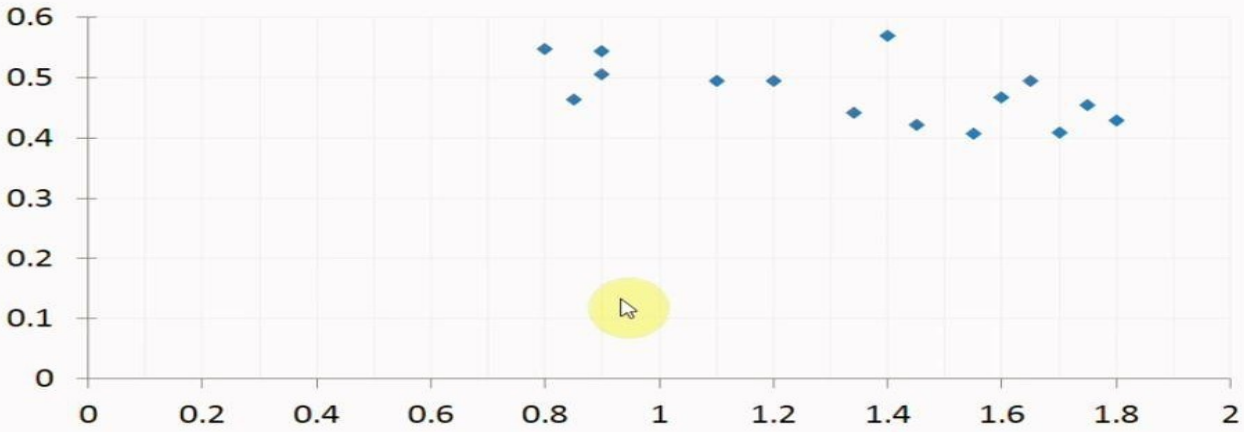
- Which dimension data appears to have more variance?
- Why do you think, we need to know where variance is more ?
- Think of examples of grade, movie success ....

Which dimension is varying more



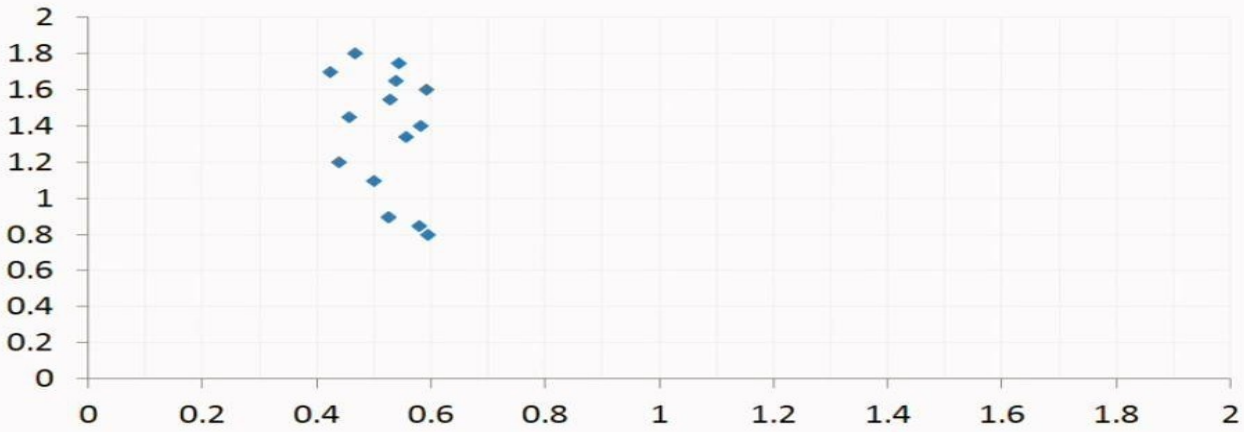
- You can draw lines perpendicular to X axis from all points.
- Where these vertical lines are touching on X axis, is called projection point.
- Now you can see the variance of projection points on X axis. i.e. The variance of data points in X dimension
- Same way by projecting on Y axis, you can see the variance in Y axis

Which dimension is varying more

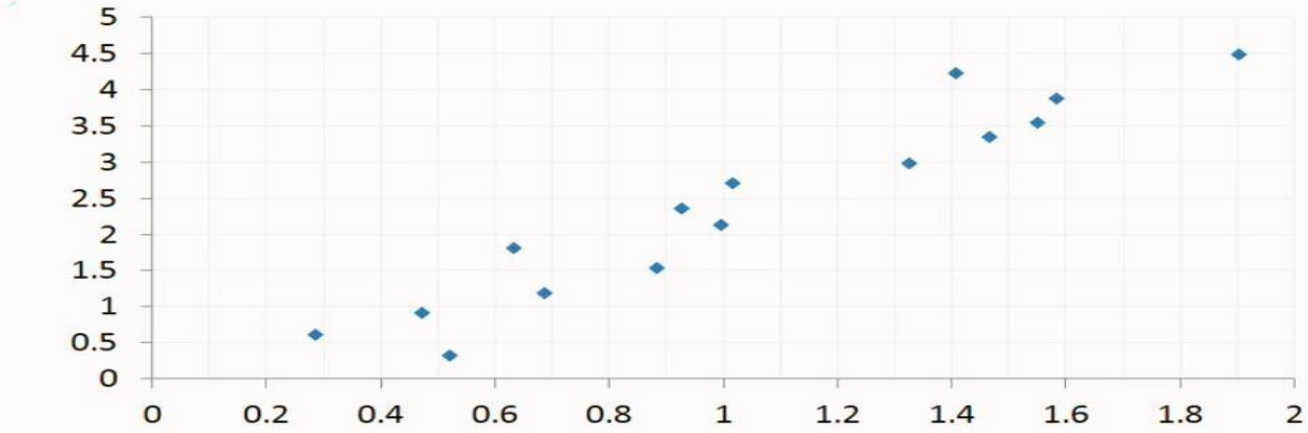


- Or you can take a look at the grid lines to get an idea
- Lots of variance in X dimension.
- Much less variance in Y dimension

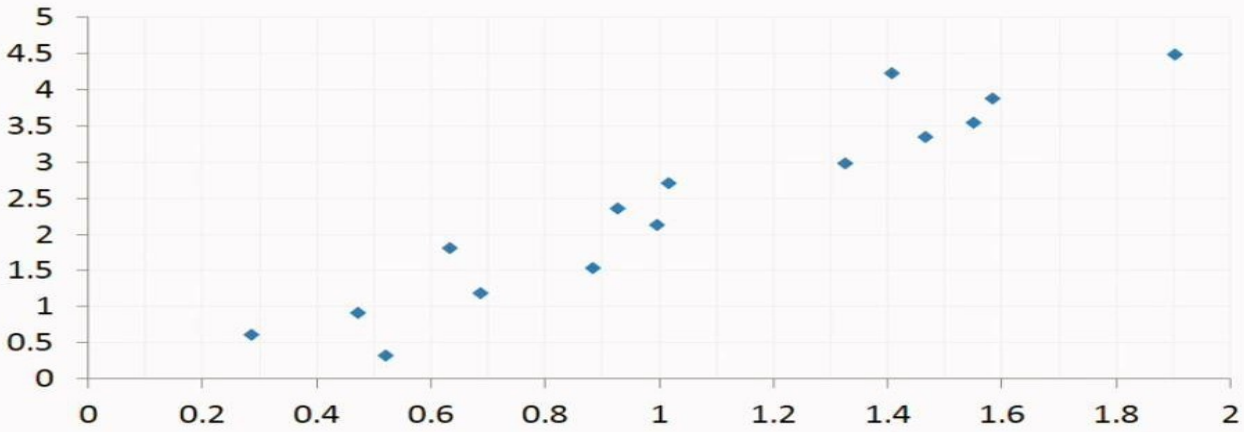
Which dimension is varying more



Which dimension is varying more

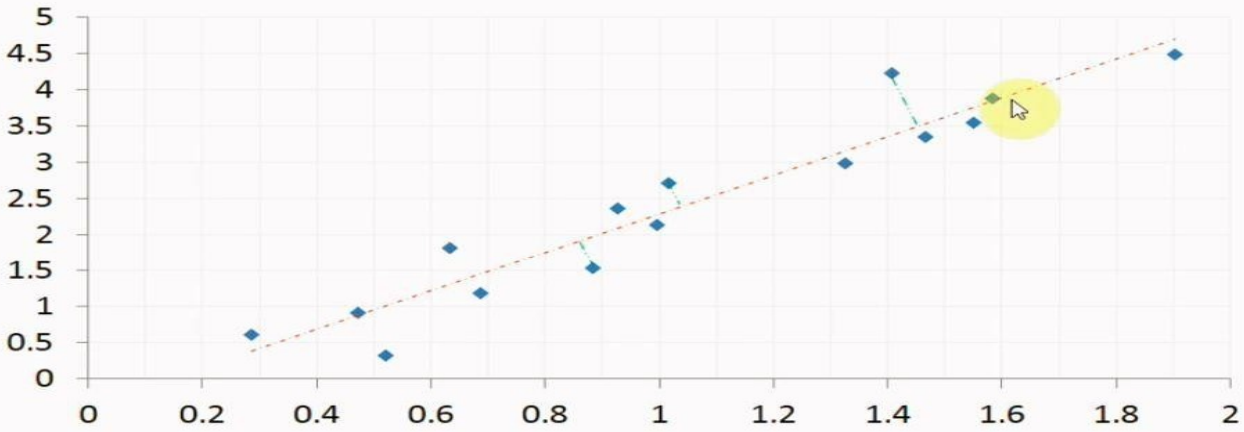


Which dimension is varying more



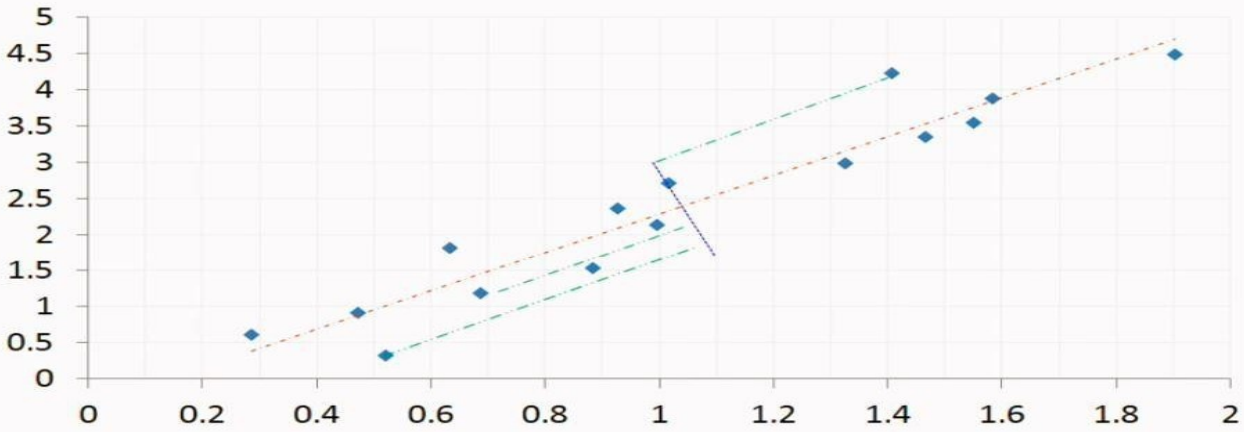
- Lots of variance in X and Y dimension.
- Can we find another direction, where variance is more than X dimension and Y dimension?

Which dimension is varying more



- If you think of, the data is varying much more in the direction of red dashed line than in the X dimension and Y dimension
- One can project on red dashed line (by drawing lines perpendicular to red dashed line) and see the variance of projection points

Which dimension is varying more



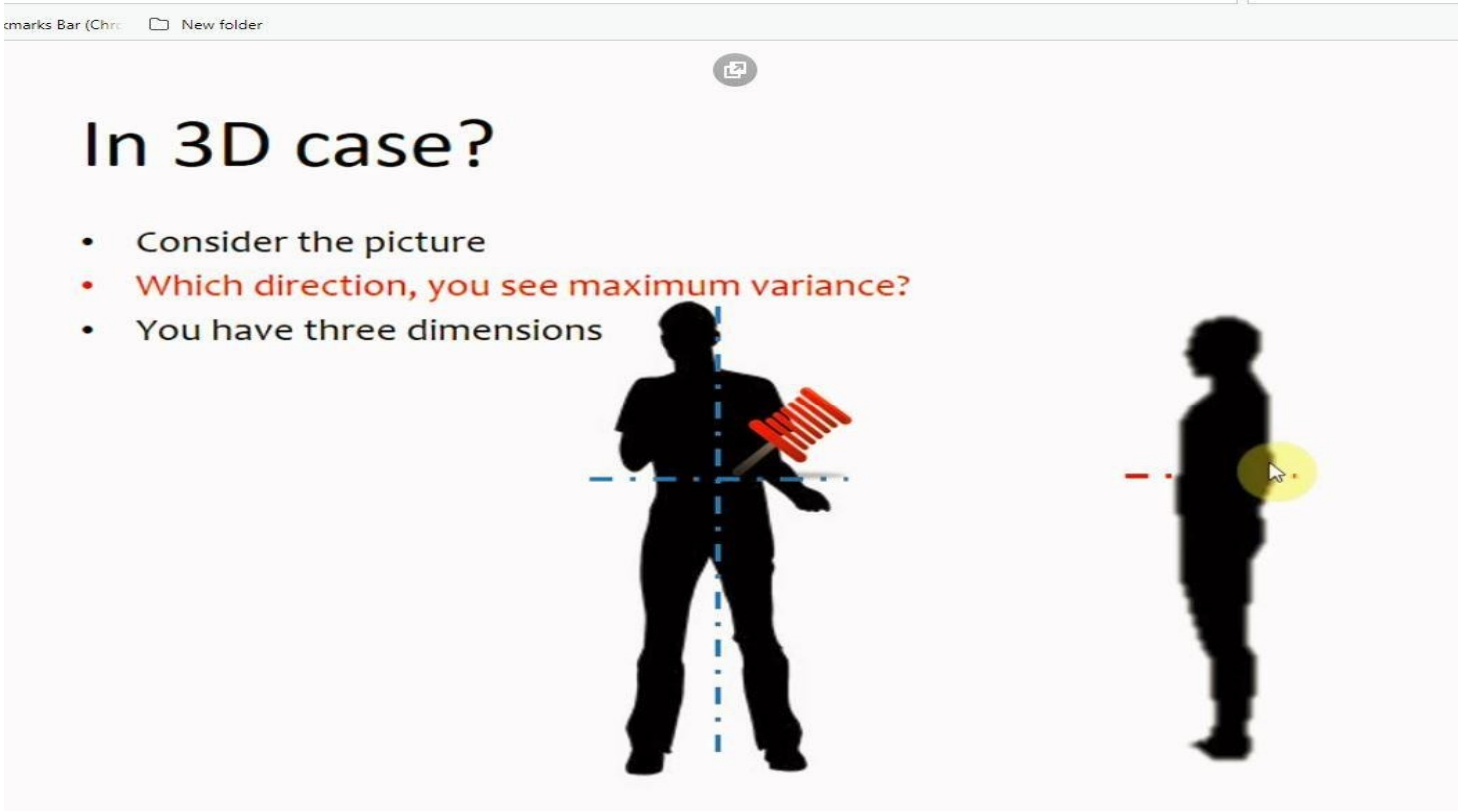
- And data is varying much less in the direction of blue dotted line, which is perpendicular to the red dashed line
- One can project on blue line and see the variance of projection points.
- This will be the variance of data in the direction of blue dotted line

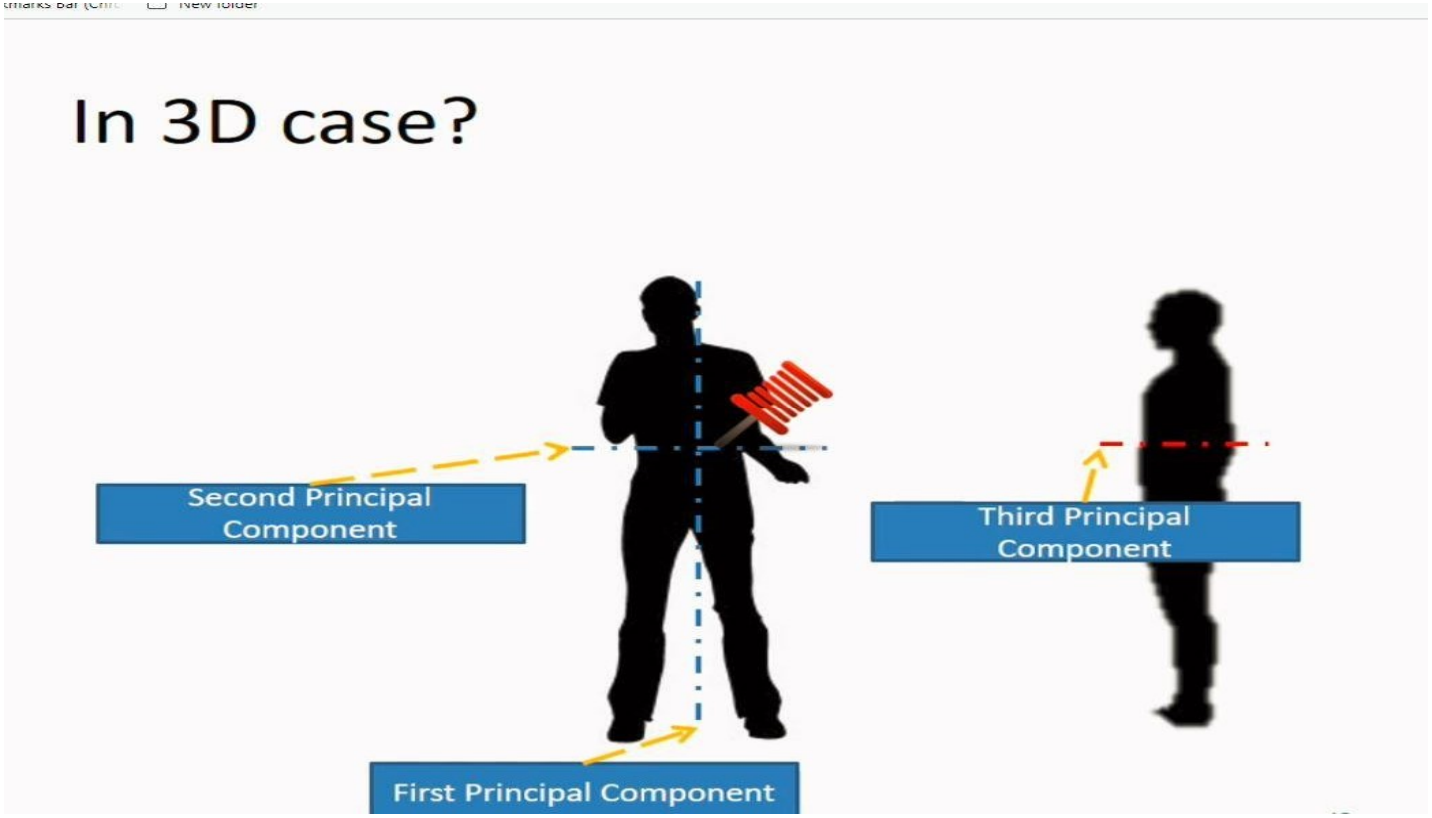
In 3D case?

- Consider the picture
- Which direction, you see maximum variance?

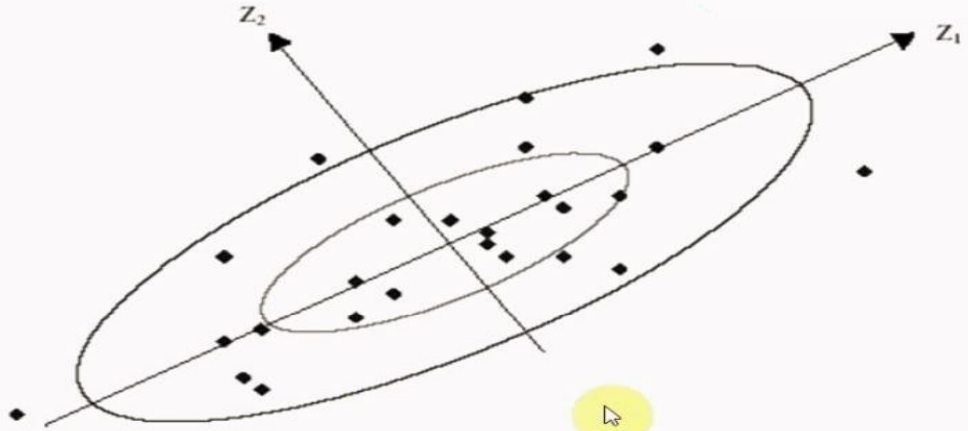
In 3D case?

- Consider the picture
- Which direction, you see maximum variance?
- You have three dimensions


marks Bar (Chr
New folder
In 3D case?
Consider the picture
Which direction, you see maximum variance?
You have three dimensions


In 3D case?
Second Principal Component
Third Principal Component
First Principal Component

PCA definition



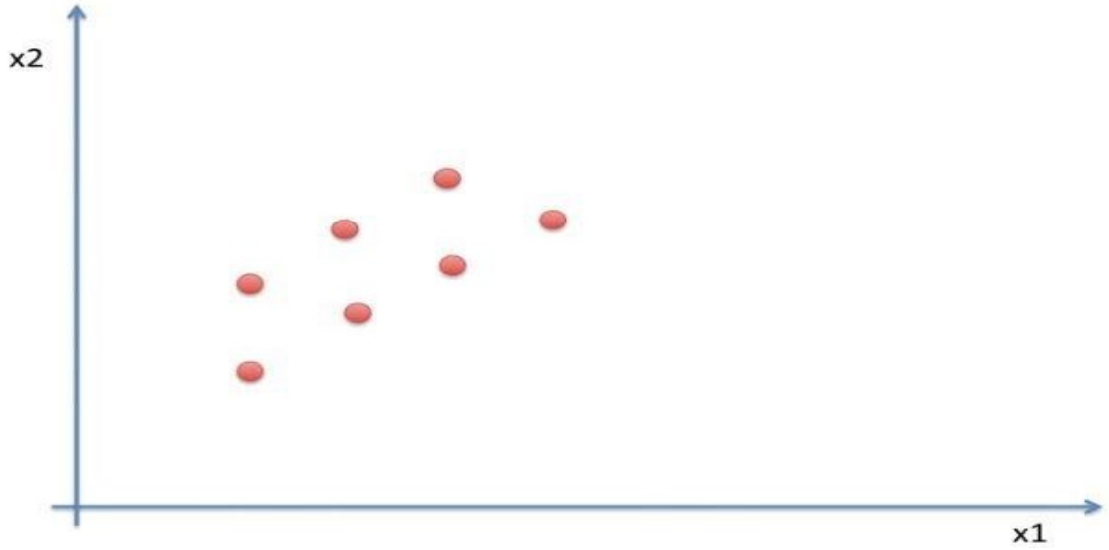
First principal component

- Among all possible axis passing through **the centre of data i.e. centre of x1 and x2**, it is the direction, on which if we draw perpendicular lines from all data points, the variance of projection points on the line z1 will be maximum.
- The sum of square of perpendicular distances of the data points from the line z1 will be minimum. (why?)

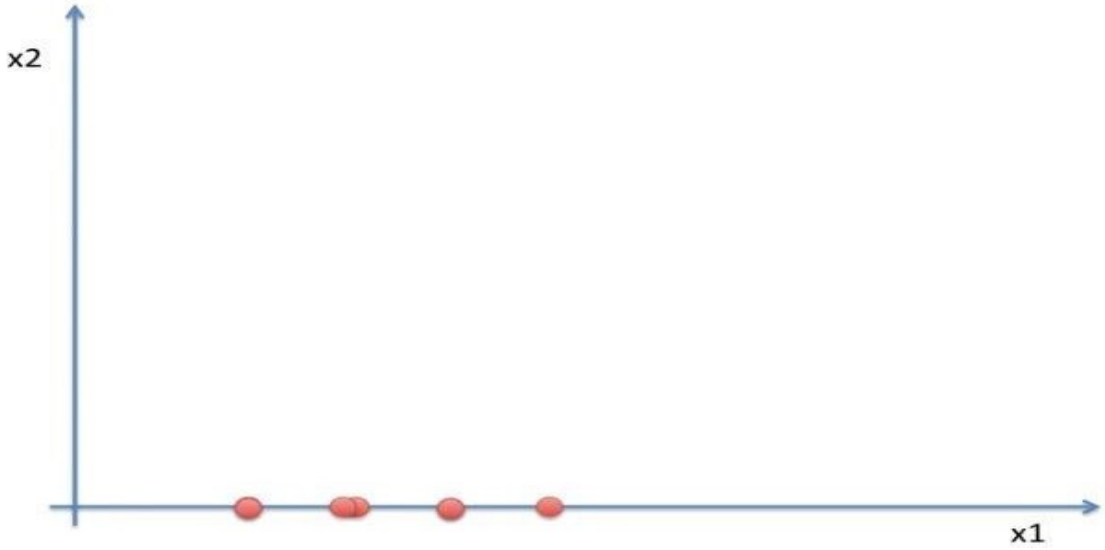
Why dimensionality reduction?

- visualization
- reduce noise
- preserve useful info in low memory
- less time complexity
- less space complexity

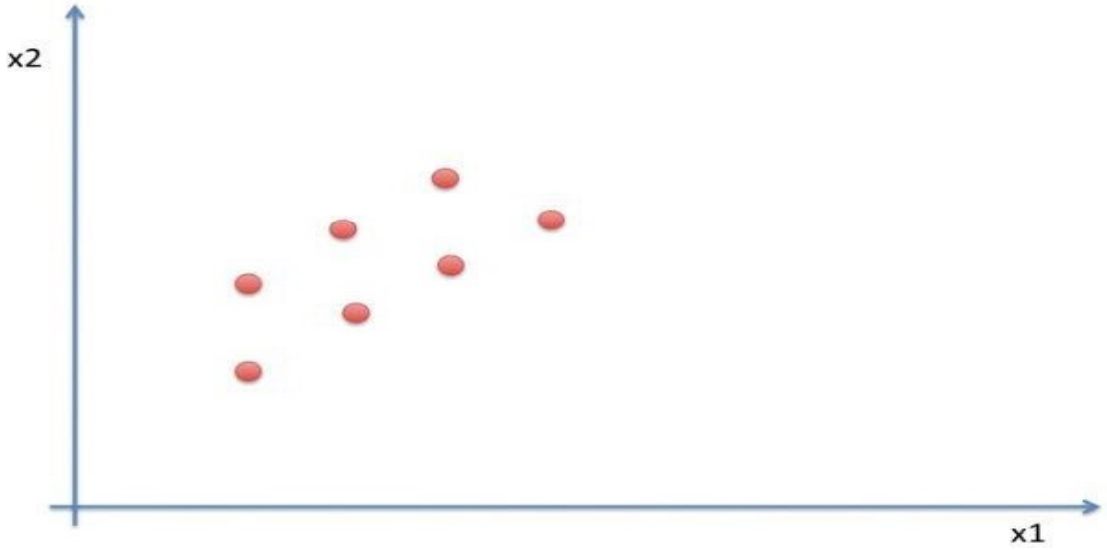
How PCA reduce dimension?



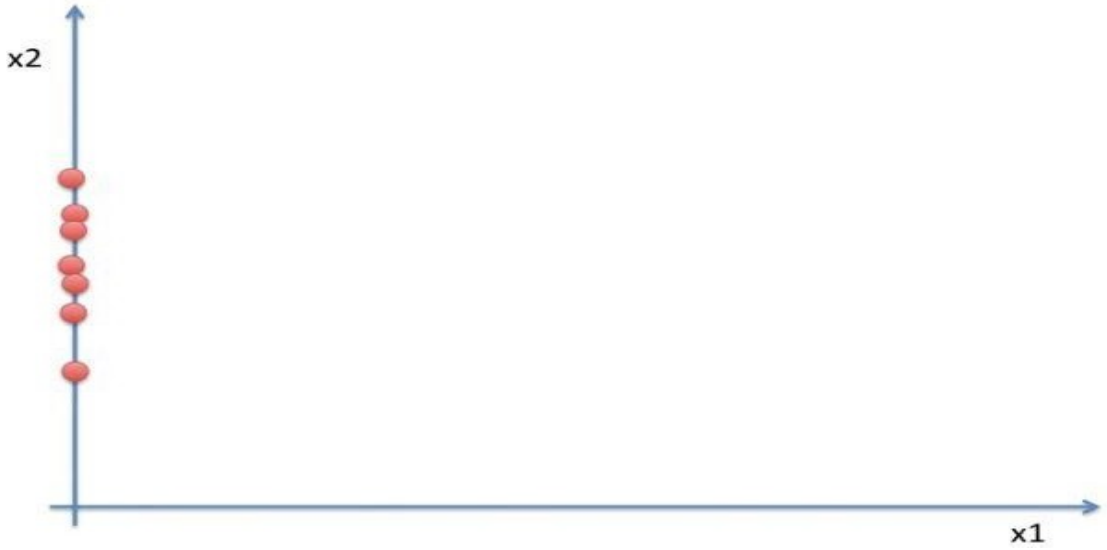
how about to use x1 axis?



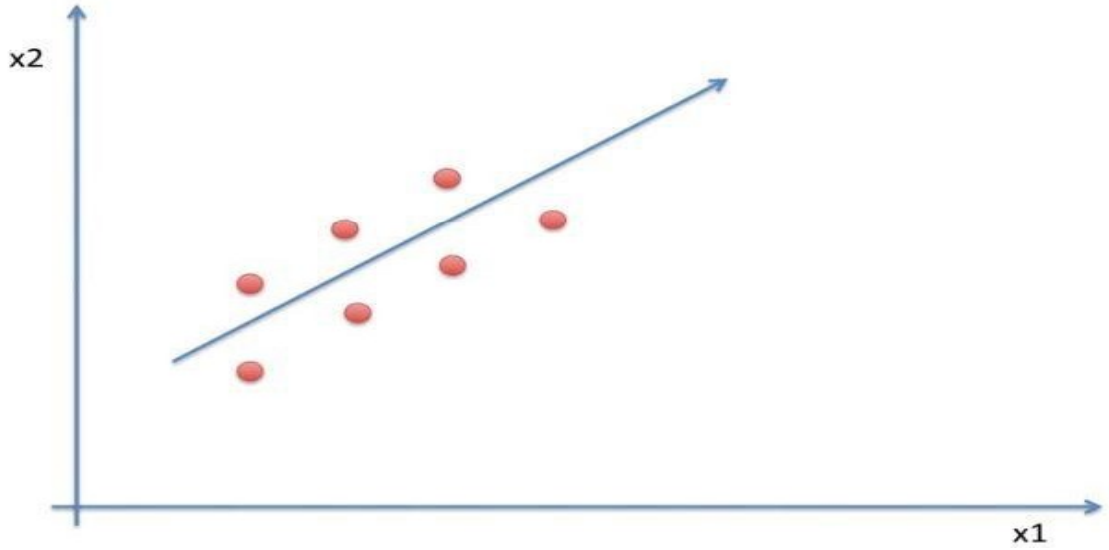
how about to use x2 axis?



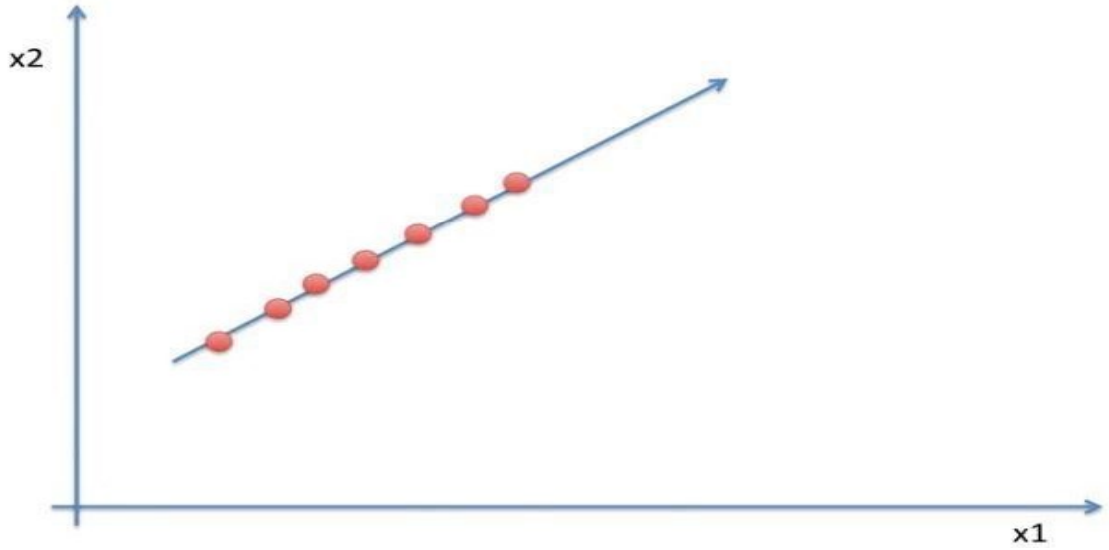
how about to use x2 axis?



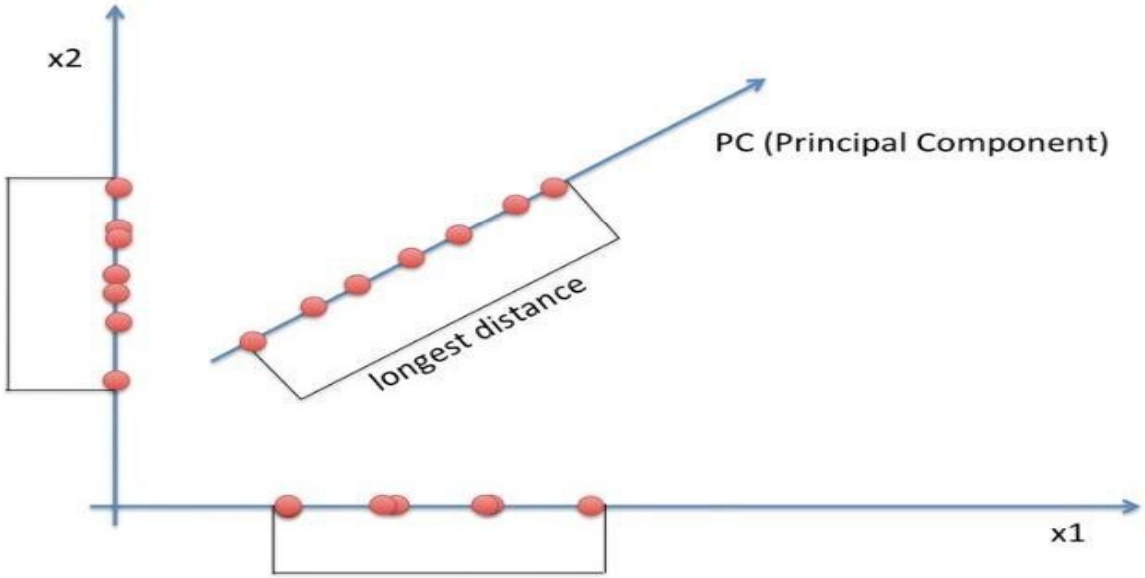
How PCA reduce dimension?



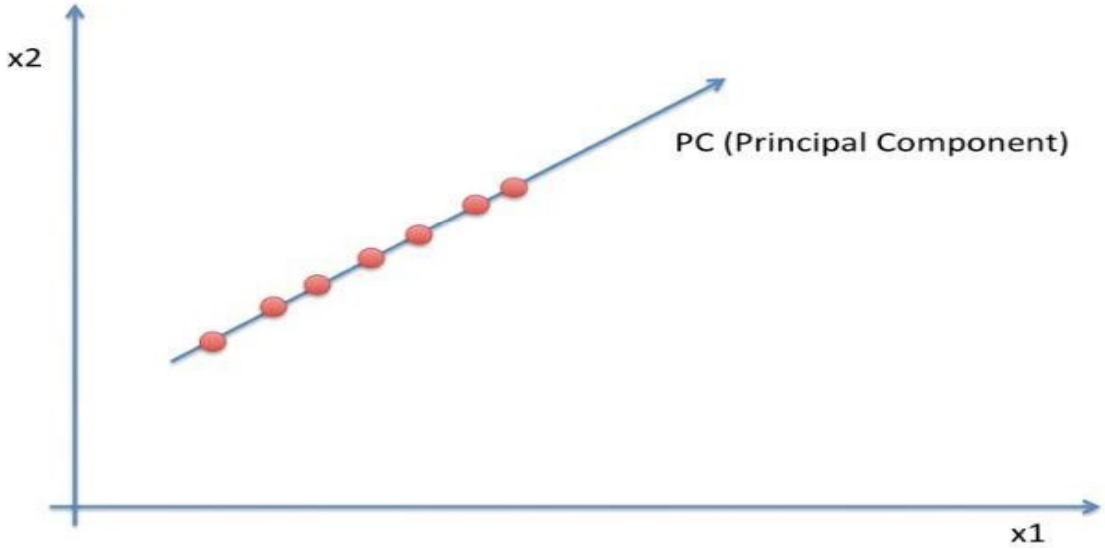
How PCA reduce dimension?



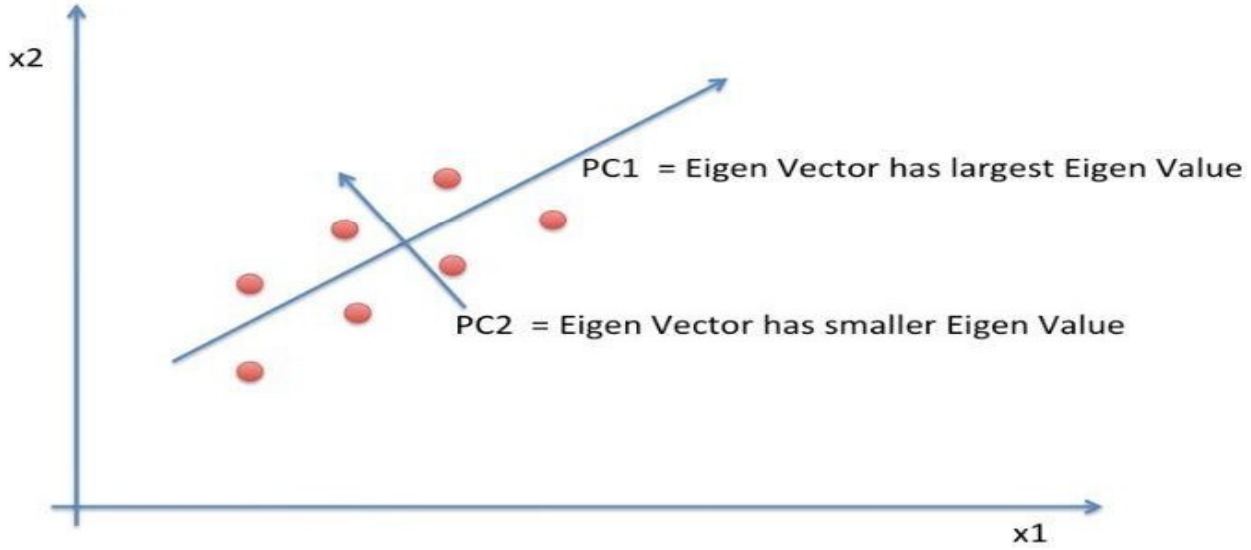
PC variance is the largest



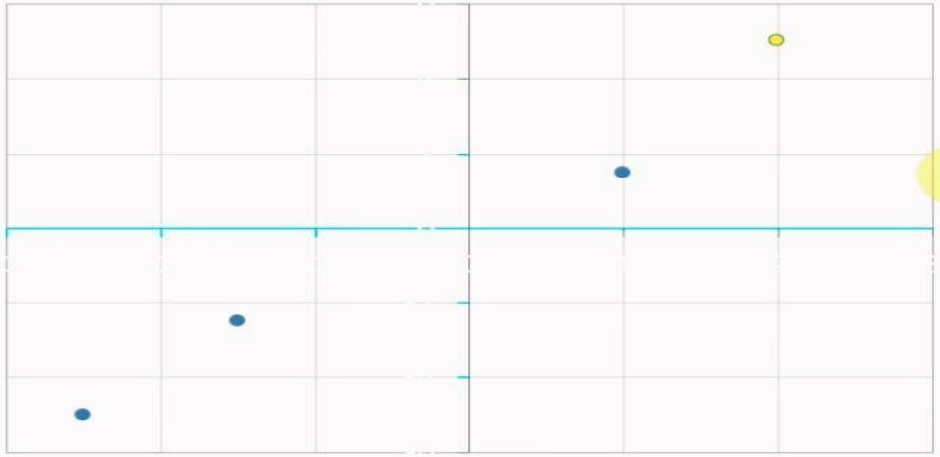
we reduced dimension to 1d in PCA algorithm



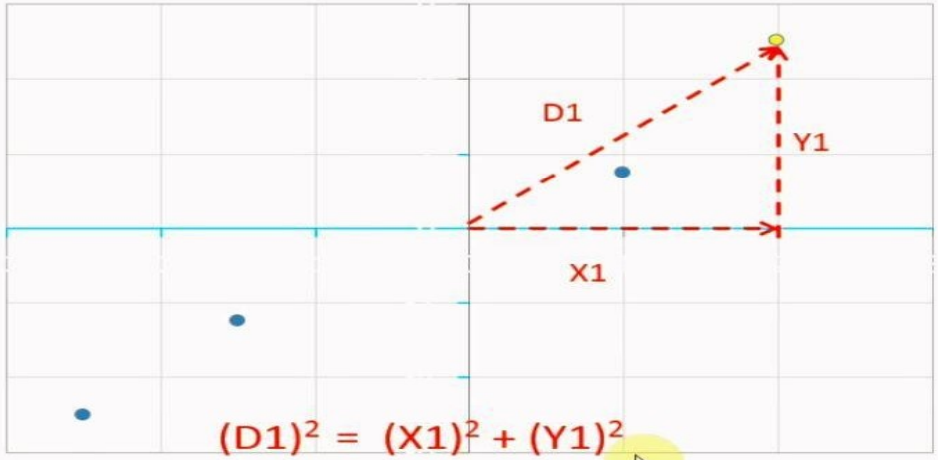
1. We select Eigen vector has largest Eigen value from covariance matrix



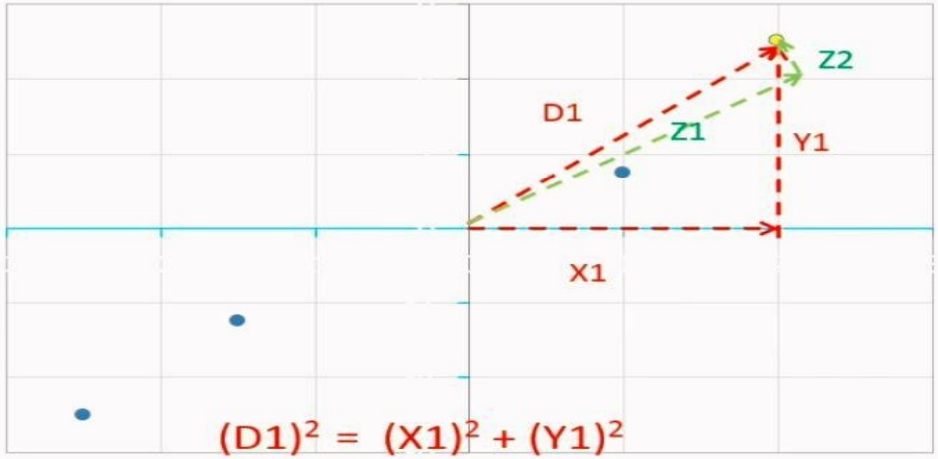
Minimum sum of squar of perpendicular distances



Minimum sum of squar of perpendicular distances



Minimum sum of squar of perpendicular distances



$(D1)^2 = (X1)^2 + (Y1)^2$

$(D1)^2 = (Z1)^2 + (Z2)^2$

Now if $(Z1)^2$ is bigger then $(Z2)^2$ has to be smaller as the sum is fixed

- N principal components for N input variables
- The goal of PCA is to find the first linear combination of input variables, which maximizes the variance of the data
- This is called first principal component (or first eigen vector) and variance explained by this component (or in other words variance of data in the direction of first principal component) has to be maximum by very design.
  - This variance is also called first eigen value.
- Then find the other linear combination, which maximizes the variance that has not been explained by first combination (i.e. residual variance)
  - This component is called second principal component and variance in this direction is called second eigen value
- And so on for third principal component etc.
- PCA gives you artificial variables through combination of input variables
- By nature sum of variance of all PCA= total variance= variance of individual variables
- Var of PC1 (Eigen Value 1) > Var of PC2 (Eigen Value 2) > …