Principal Component Analysis (PCA)

PCA

- Unsupervised Machine Learning technique to reduce Dimensionality – objective of PCA

- Transforms columns of the dataset into a new set of features
  - These new features are called ***Principal Components***

PCA

- Unsupervised Machine Learning technique to reduce Dimensionality

- PCA aims to extract ***p*** features from the total ***f*** features ($p \leq m$) of a dataset
  - Which feature is more valuable to cluster the data
  - To explain the most variance of the dataset
  - Regardless of dependent variable ($\hat{\mathbf{y}}$)
  - Visualisation becomes better and more informative

- PCA performed on a correlation matrix
  - Numeric Dataset
  - Standardized dataset *

Typical problems with a huge dataset

- Unwanted features
- Features may exhibit multicollinearity
- Features may exhibit singularity
- Indecisiveness + may lead to building a bad model *

+

  - Right set of features
  - Right algorithm

*

  - Overfit
  - Underfit
  - Poor Accuracy

- Formula for the Number of (scatter) plots on a given dataset = **p(p-1) / 2**

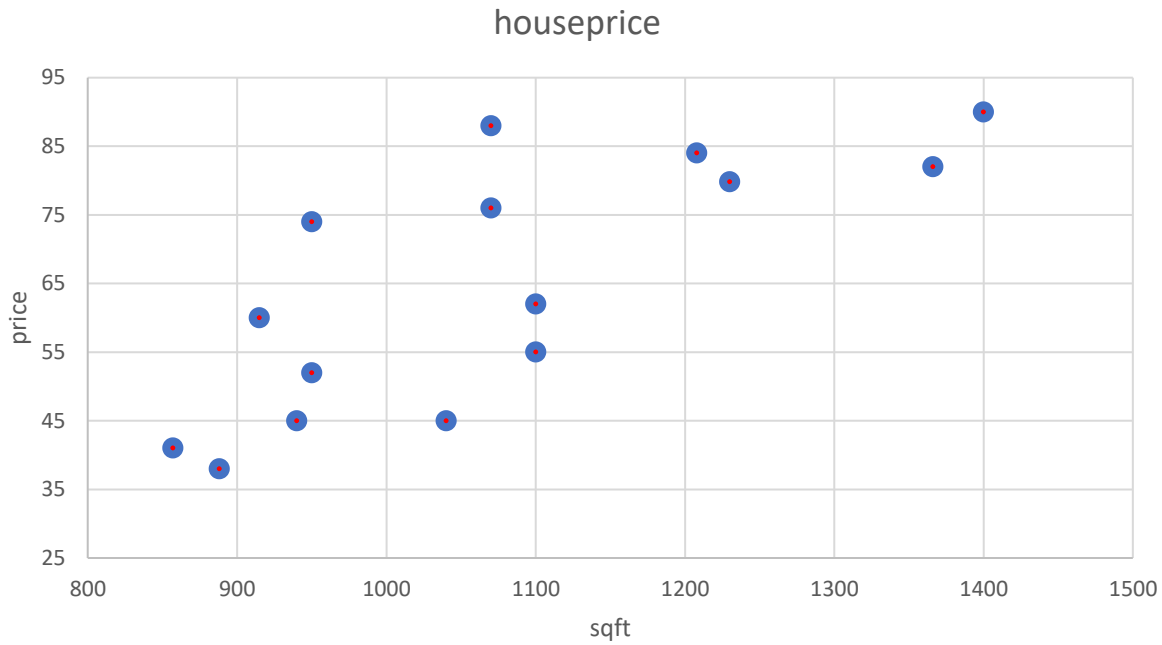
p = number of features / predictors / independent variables

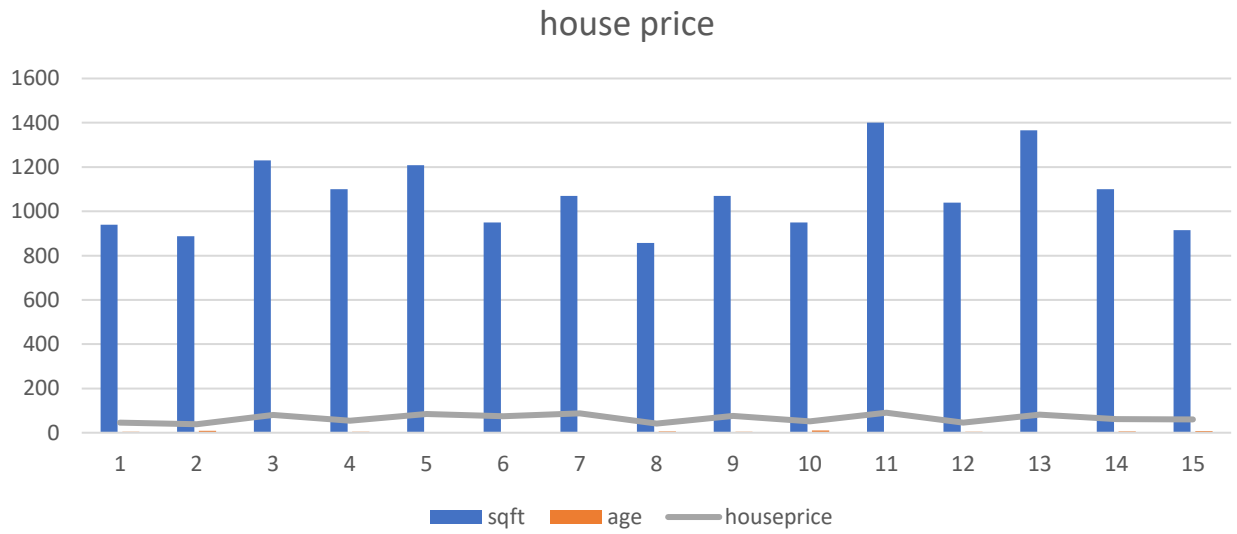
- Greater the value of p (i.e. more features), greater will the plots
  - Difficult and tedious to perform analysis

Principal Components

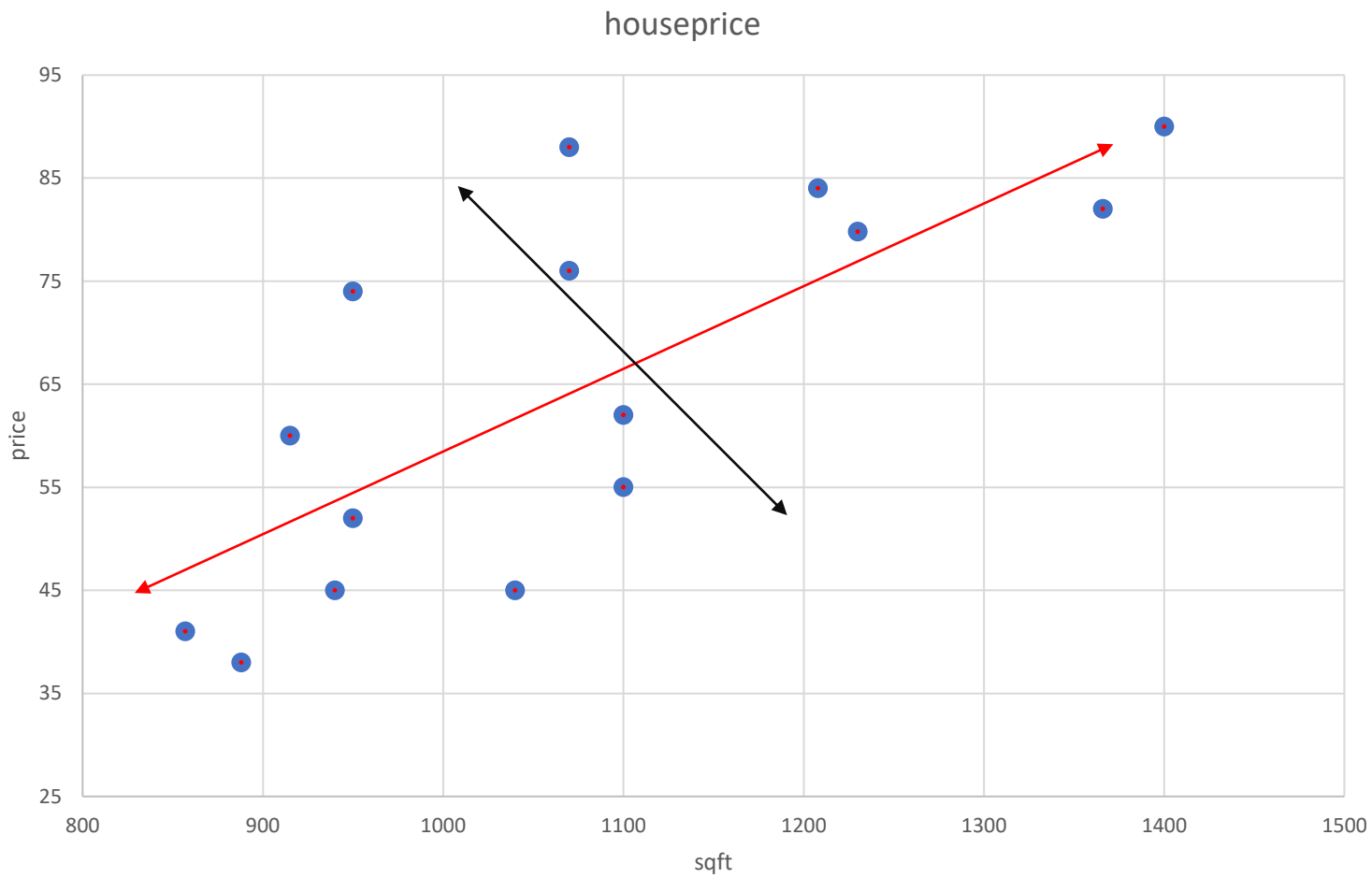
- A Principal Component is a linear combination of normalized features from the original dataset
- PC's can be
  - **Z1**: First Principal Component
    - ✓ Captures the maximum variance
    - ✓ Captures the highest variability
  - Z2: Second Principal Component
    - ✓ Captures the remaining variance
    - ✓ Uncorrelated to Z1
  - **Z3, Z4 ….**
- The number of PC's that can be constructed for a **n**x**p** dataset is:
  - ✓ **min(n-1, p)**

sqft	age	houseprice
940	5	45
888	9	38
1230	2	79.8
1100	4	55
1208	1	84
950	3	74
1070	2	88
857	6	41
1070	4	76
950	10	52
1400	1	90
1040	4	45
1366	3	82
1100	6	62
915	8	60

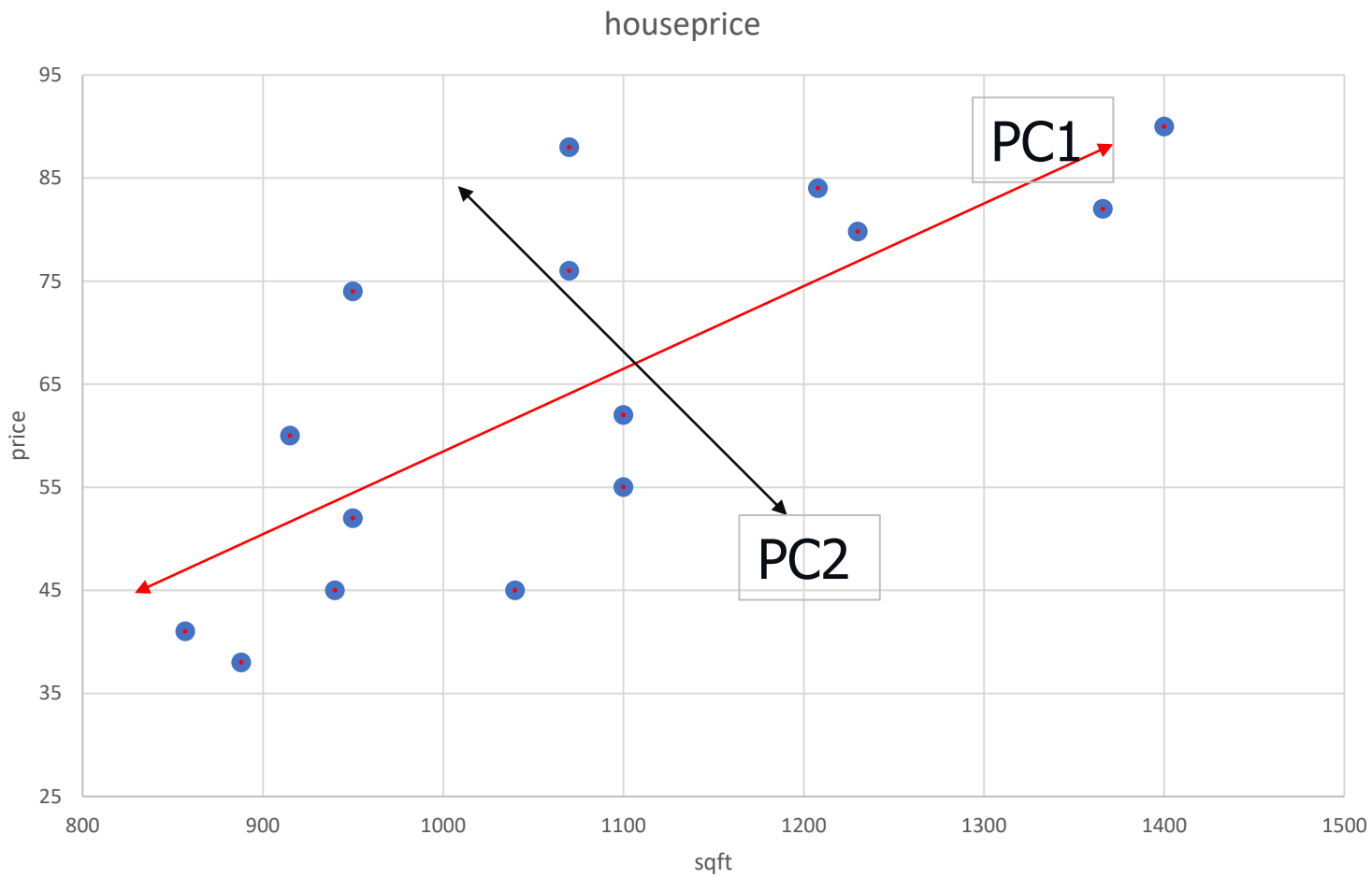




- With increase in dimensions, the chart becomes more difficult to plot
- Are all dimensions relevant / important ?
- 2-dimensional vs 3-dimensional movies
  - ➢ less information loss in 2-D
- PCA flattens dimensions



- Maximum variation
  - ➢ Left – Right (Red line)
- 2nd most Variation
  - ➢ Top-bottom (Black line)

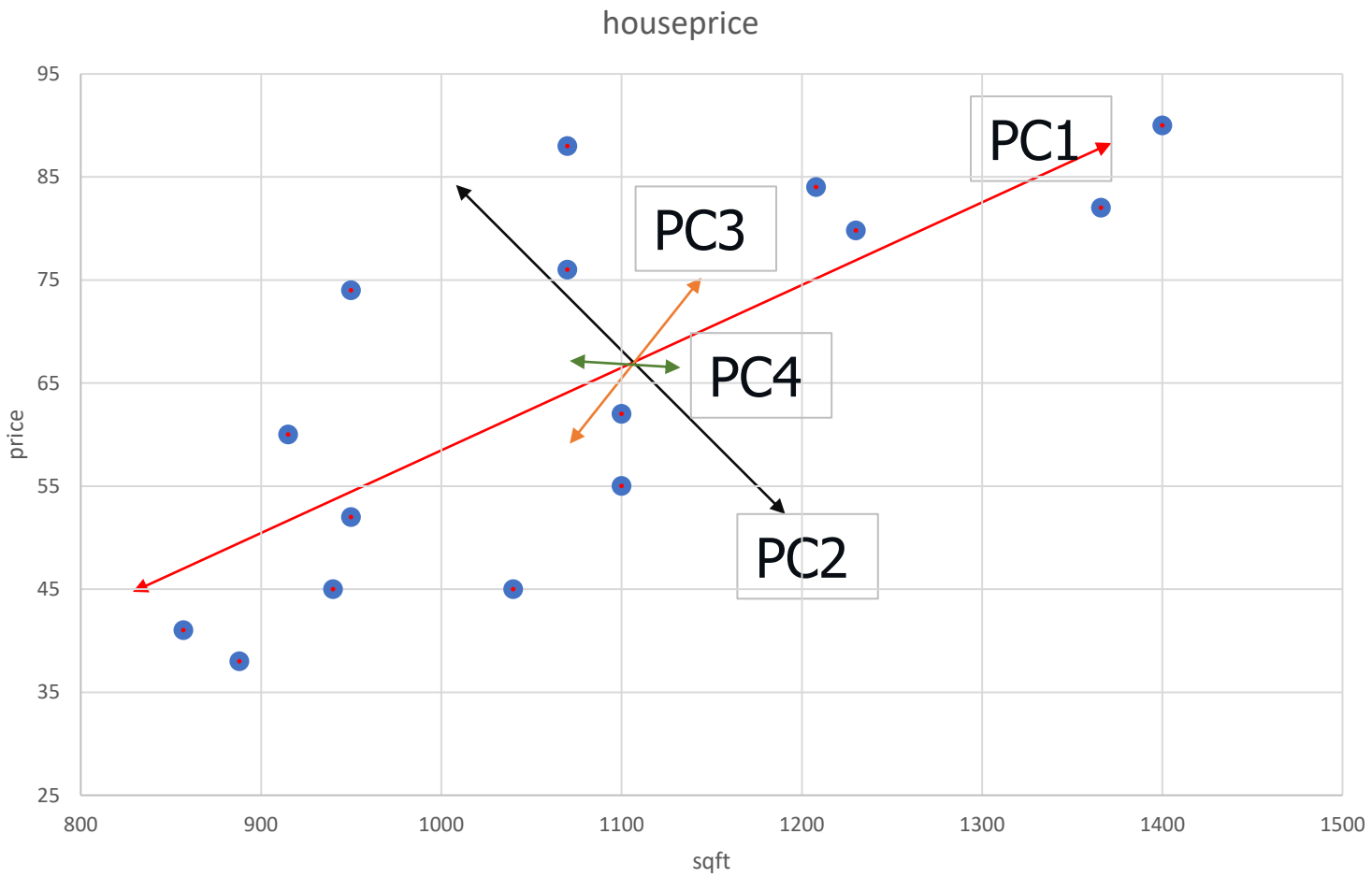


- Maximum variation
  - Left – Right (Red line)
  - PC1
- 2nd most Variation
  - Top-bottom (Black line)
  - PC2

Maximum variation is captured with these 2 lines

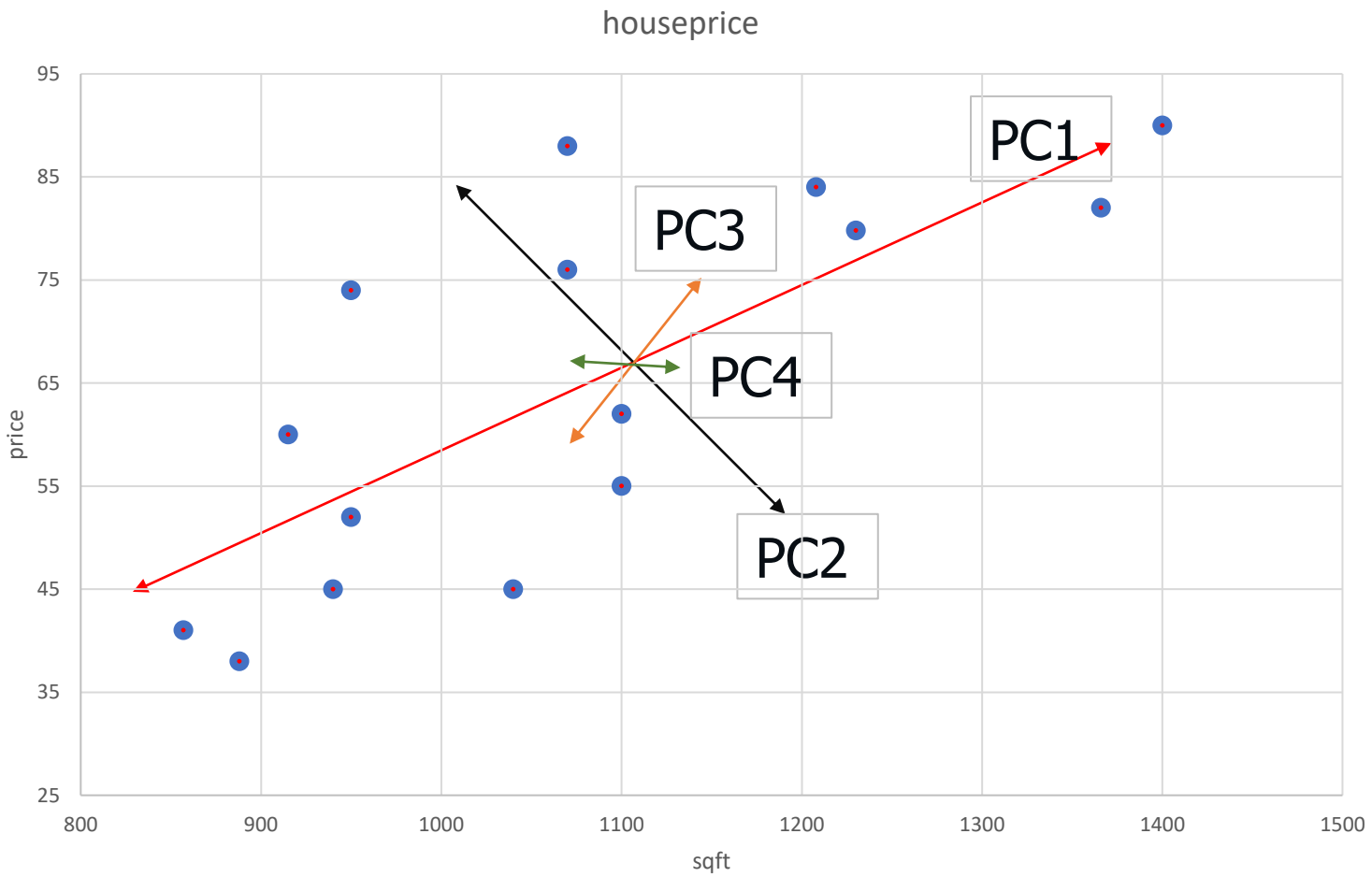
No real need for a diagonal line to capture more variation

When there are multiple dimensions



- PC1
  - ➢ Maximum variation
- PC2
  - ➢ 2nd most variation
- PC3
  - ➢ 3rd most variation
- PC4
  - ➢ 4th most variation

Principal Components – finer points



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.

No other component can have variability higher than first principal component.

If two components are uncorrelated, their directions should be orthogonal (90 degrees)

The principal components are supplied with normalized version of original predictors. This is because, the original predictors may have different scales.

How to perform PCA

- Calculate the mean value of feature (X)
- Calculate the Covariance matrix
- Find the Eigen Value
- Find the Eigen Vector

Identify the Principal Component

- Eigen vector having the highest Eigen value is the First Principal Component

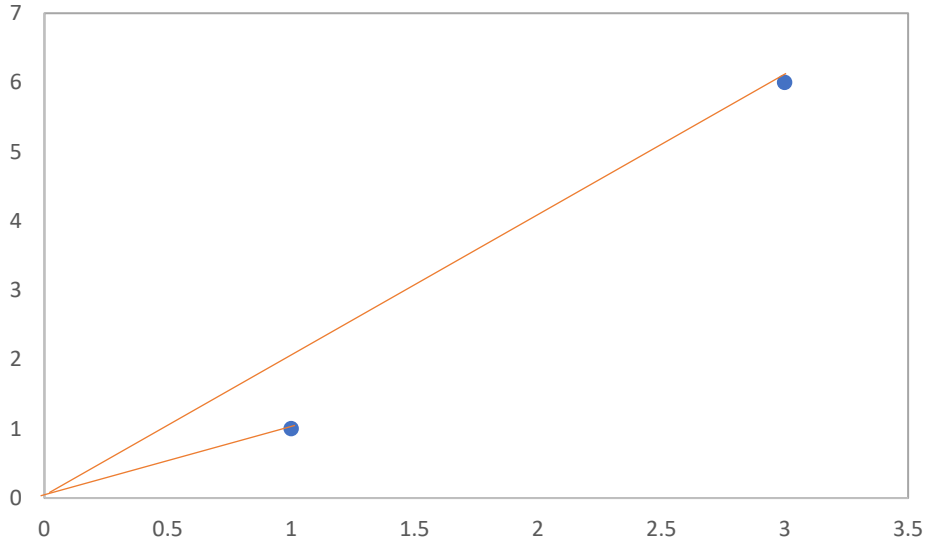
Eigen Vector

An eigen vector is a vector that scales itself, without undergoing any rotation

Eigen Value

Scaling factor

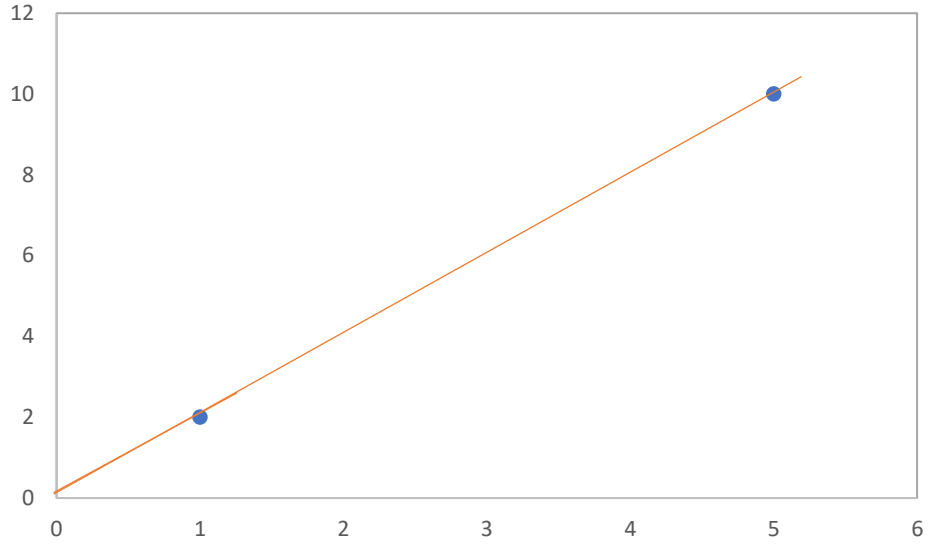
A		x		Ax
1	2	1	=	3
2	4	1		6



Scaled, Rotated

A		x		Ax
1	2	1	=	5
2	4	2		10

EV = 5



Scaled, Not Rotated

$$Ax = \begin{bmatrix} 5 \\ 10 \end{bmatrix}$$

$$Ax = 5\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$Ax = 5x$$

$$Ax = \lambda x$$

X : Eigen vector
λ : Eigen Value

Given

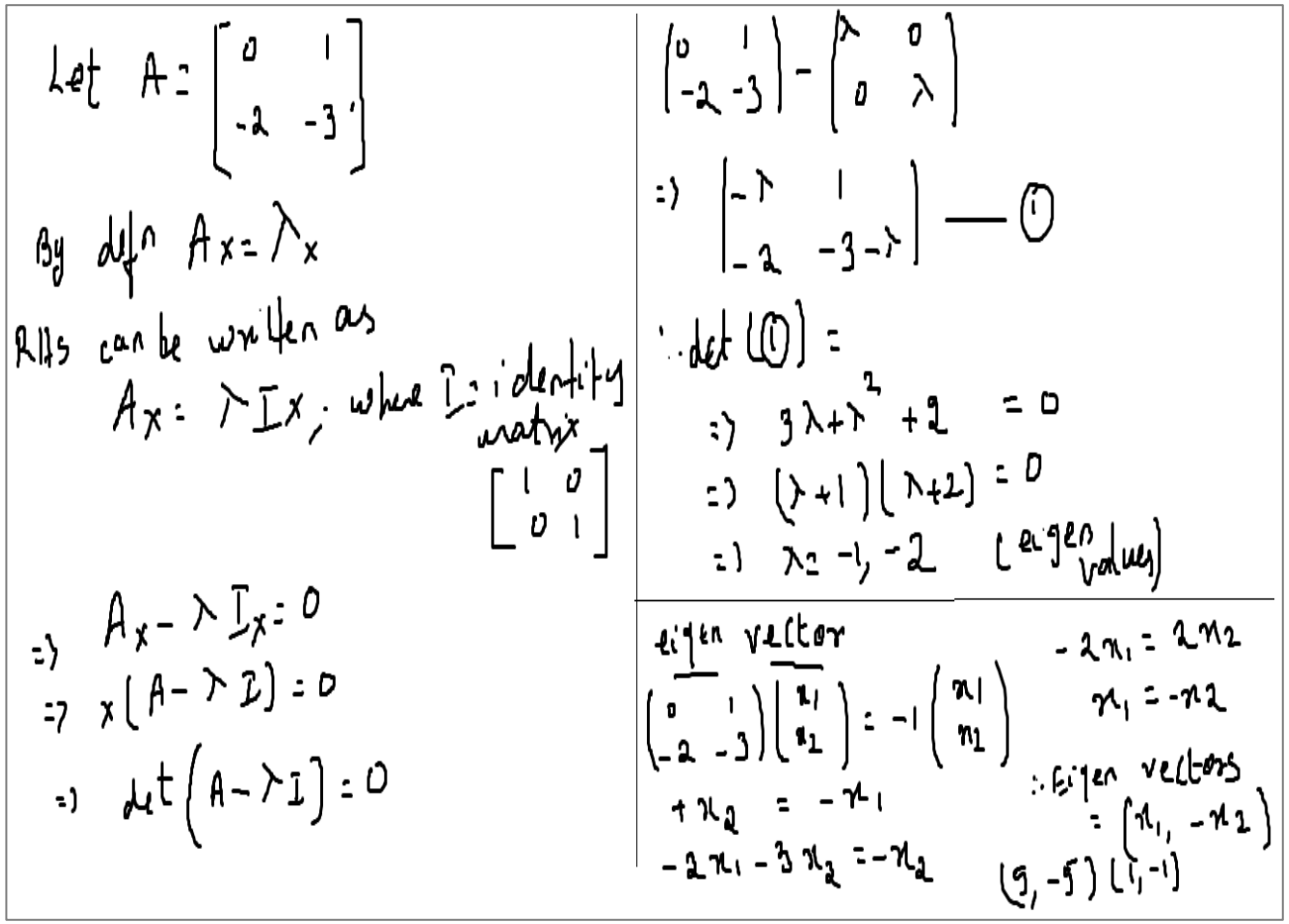
- Any square matrix A
- Any vector x <> 0

Such that Ax = λx where λ is an integer

λ is the Eigen Value
x is the Eigen Vector

How to determine the Eigen value and the Eigen Vector for a given square matrix

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$



Scree Plot

- **Graphical representation of percentages of variation of every Principal Component**