Principal Component Analysis (PCA)

Qi Zeng

What is PCA?

PCA is a dimension reduction method.

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.

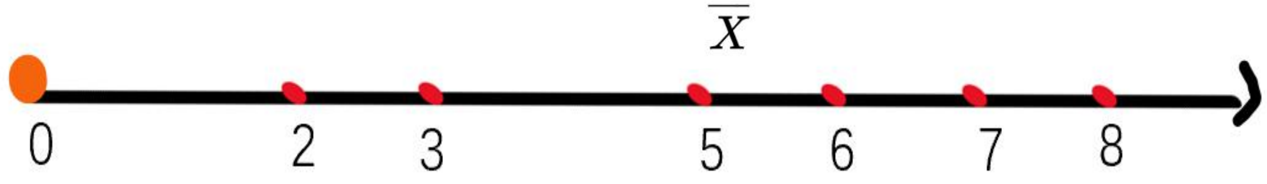
How it is used?

One dimensional:

X
2
5
3
7
8
6

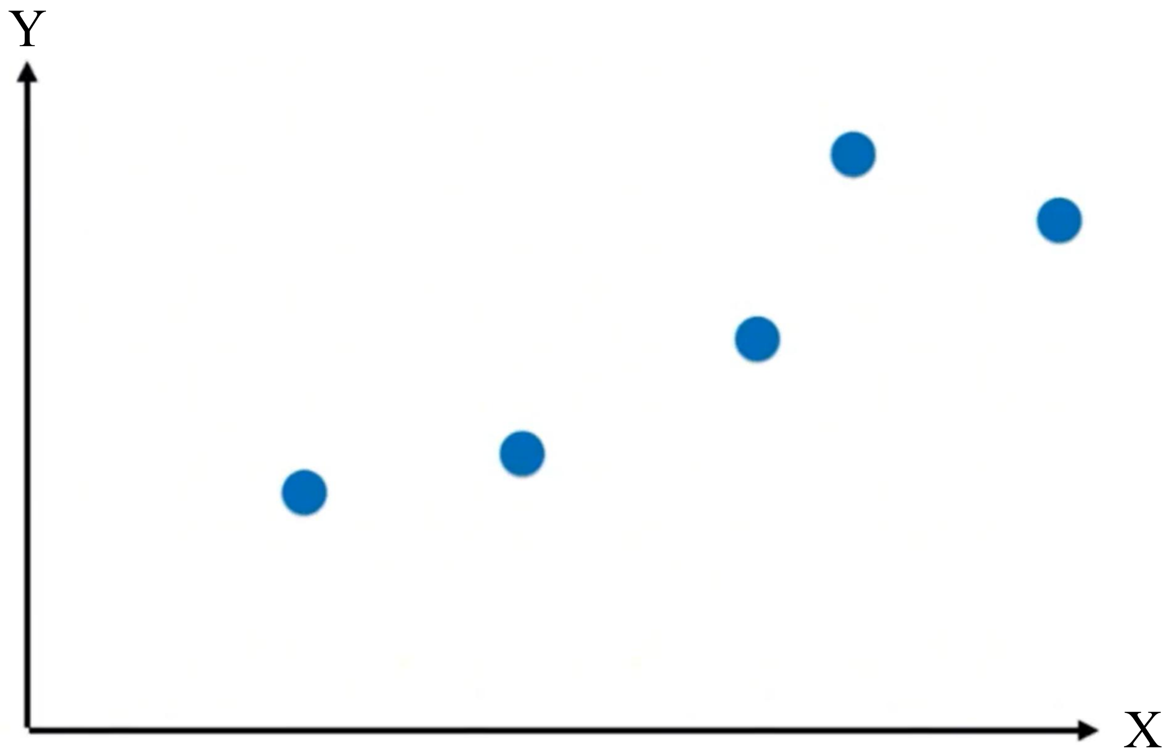
$$\overline{X} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = 5.17$$

$$Var(X) = \frac{\sum(X - \overline{X})^2}{n-1} = 4.47$$

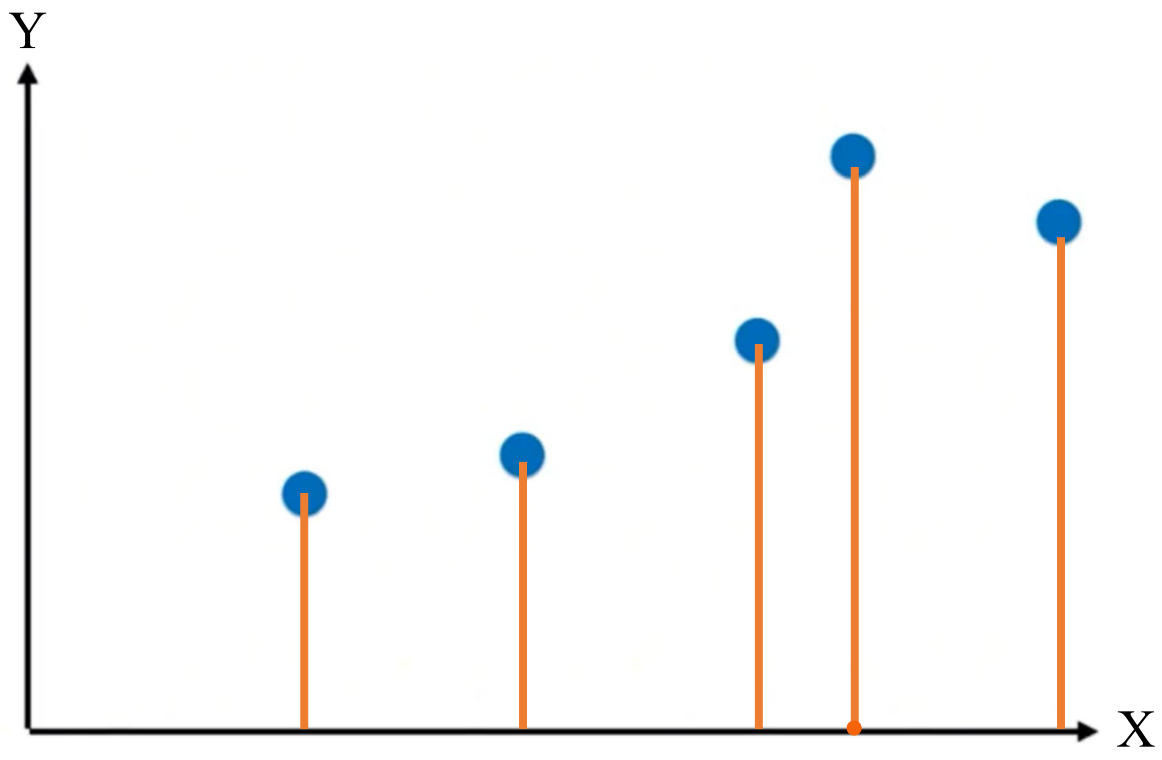


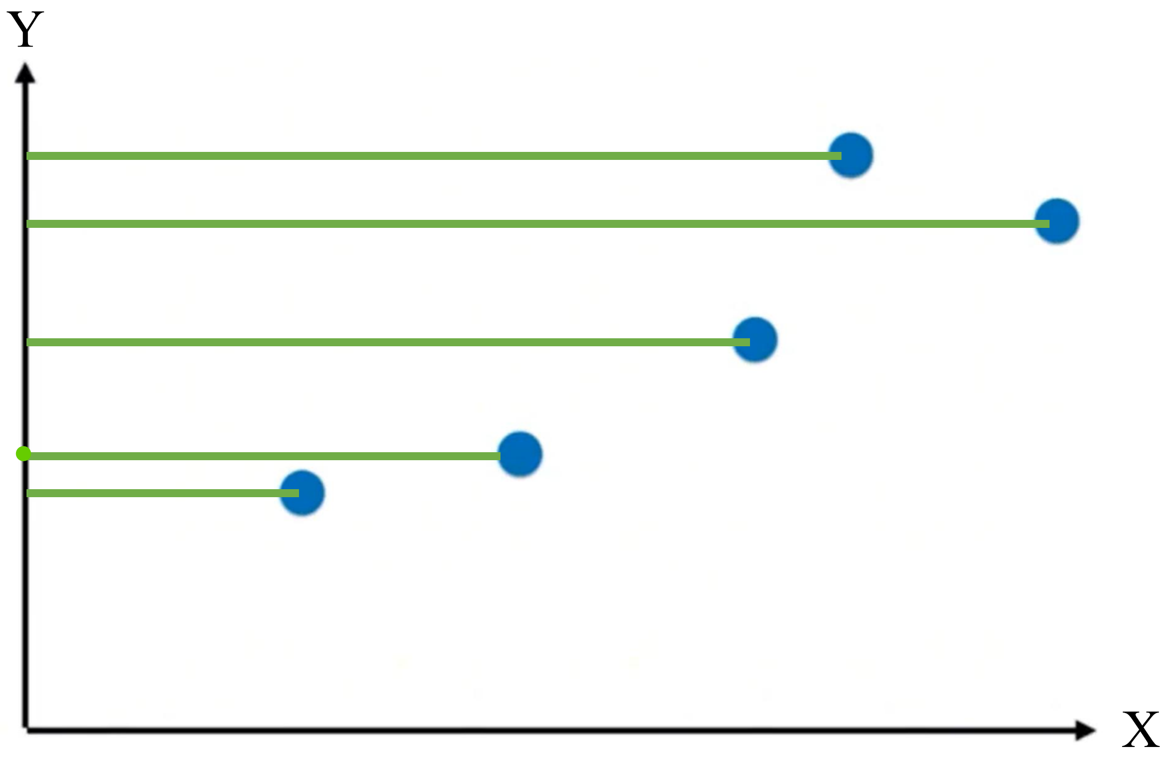
X(after demean)
-3.17
-0.17
-2.17
1.38
2.38
0.38

Two dimensional:



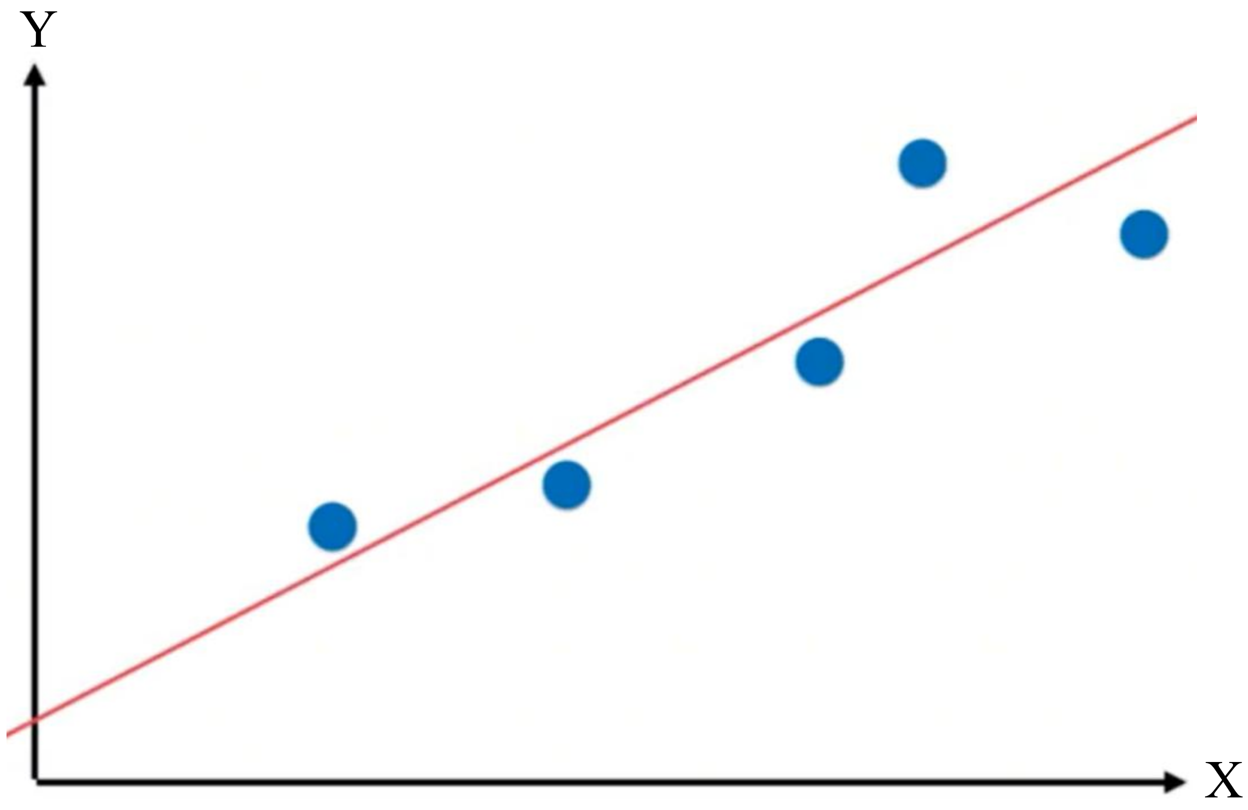
Two dimensional:





Question：
Which is better and why?
Is there a better way than these two projection?   PCA

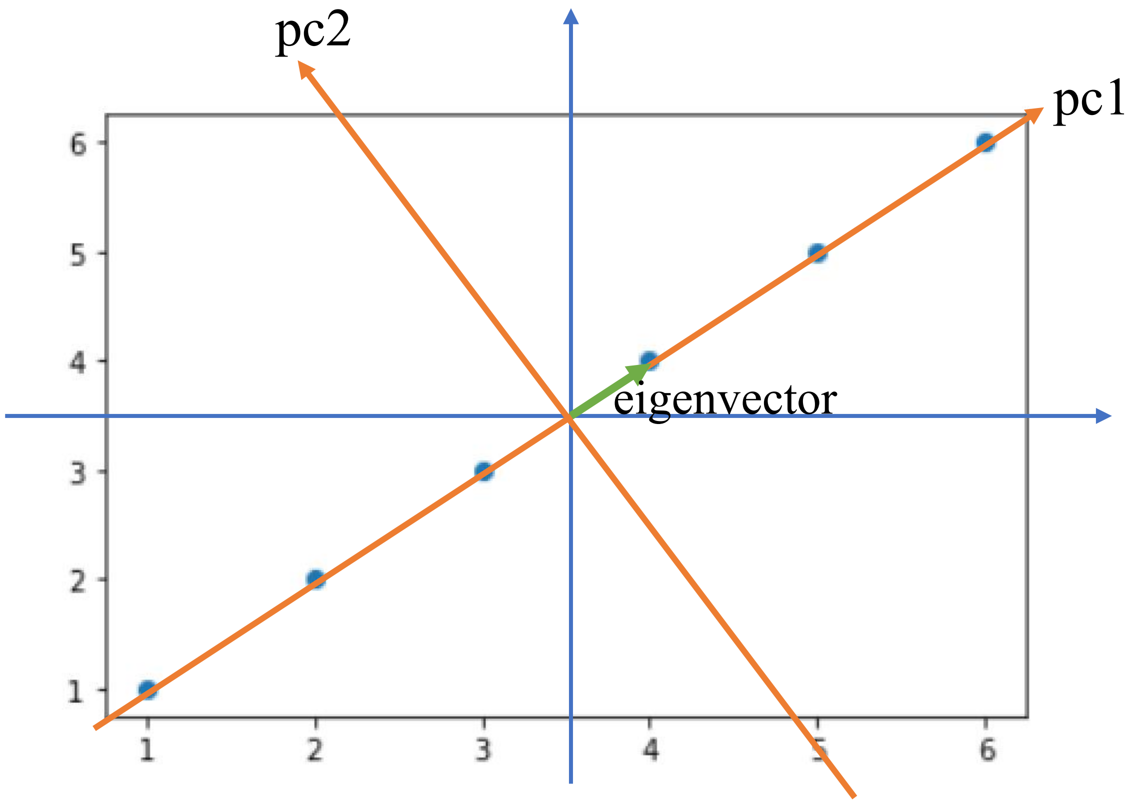
Two dimensional:



Two dimensional:

A special example

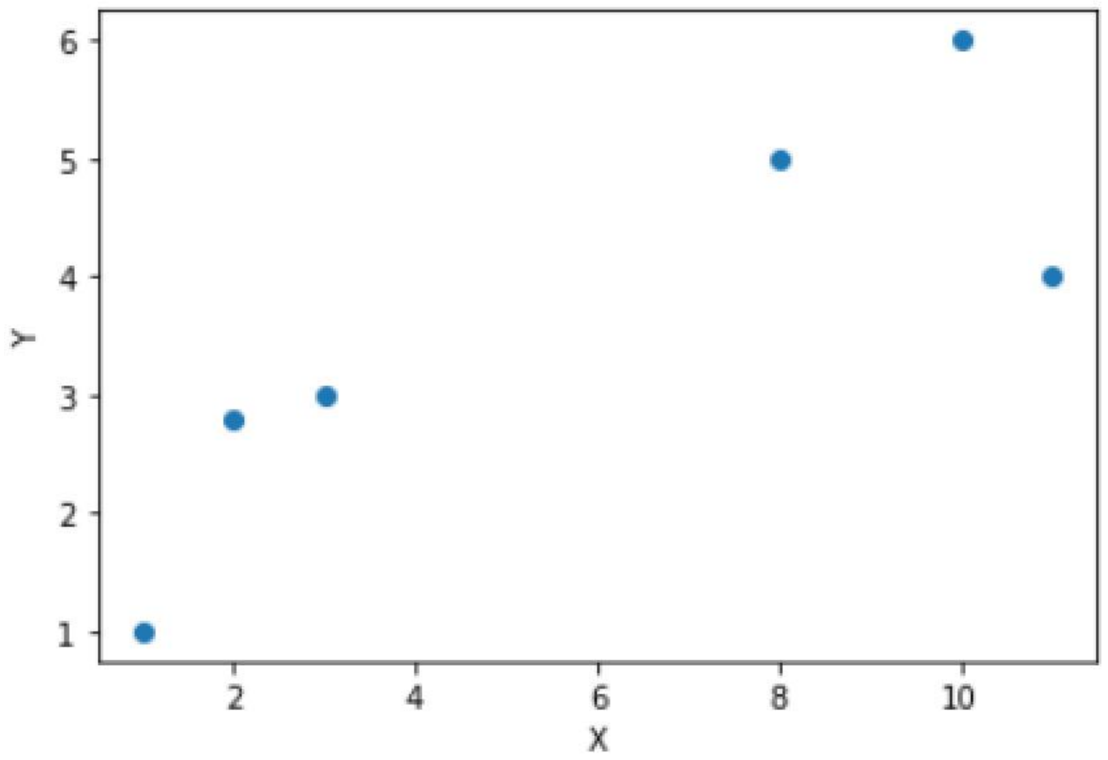
```
X=[1,2,3,4,5,6]
Y=[1,2,3,4,5,6]
```



New:
X=[-2.5 -1.5 -0.5 0.5 1.5 2.5]
Y=[0 0 0 0 0 0]

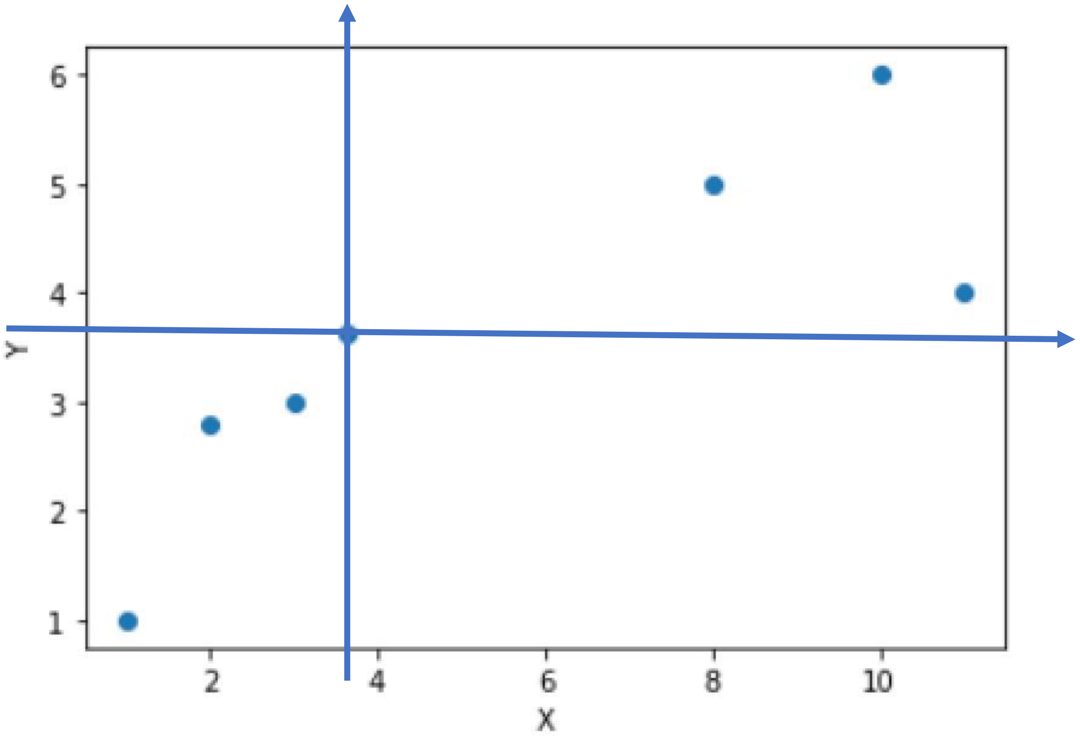
Two dimensional:

X	Y
a1	b1
a2	b2
a3	b3
a4	b4
a5	b5
a6	b6



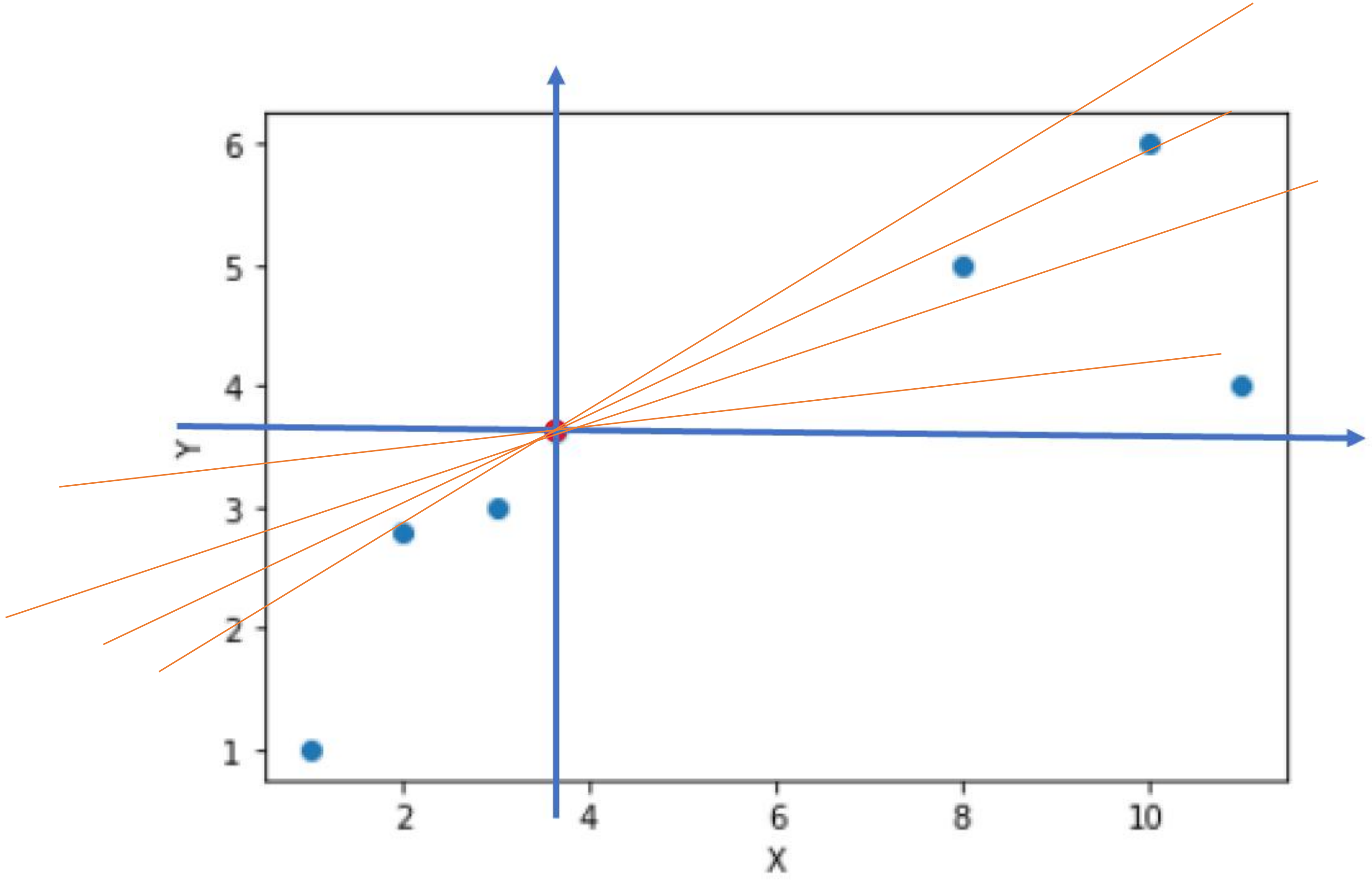
First demean:

Calculate the mean value of X and Y, and this is the center of the data.



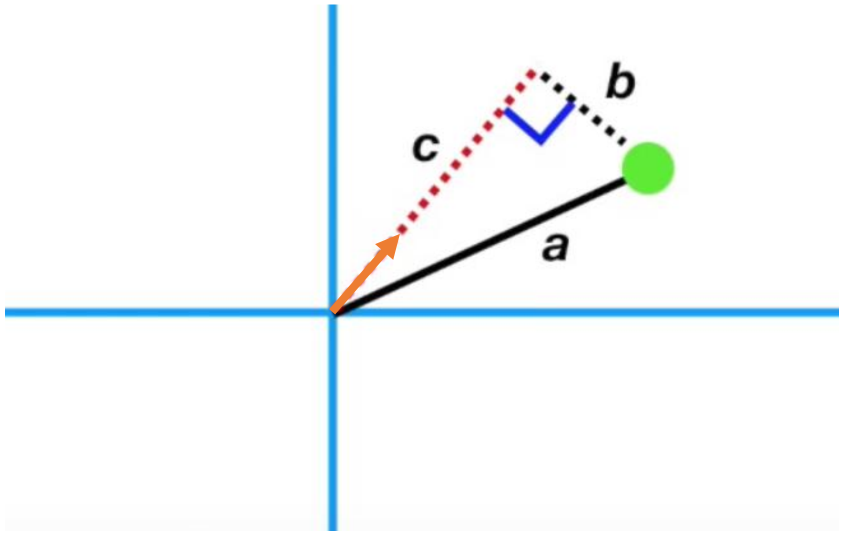
Second find a maximum value:

Cross the origin, we can draw many lines, but which is the one we want?



Second find a **maximum value:**

Take one point for example.



$$a^2 = b^2 + c^2$$

Minimize the distance to the red line,
maximize the distance from the projected point to the origin.

$$\sum_{i=1}^{m} b_i^2$$ minimum

How can do it?

$$\sum_{i=1}^{m} c_i^2$$ maximum

Second find a **maximum value:**

Linear algebra can help us.

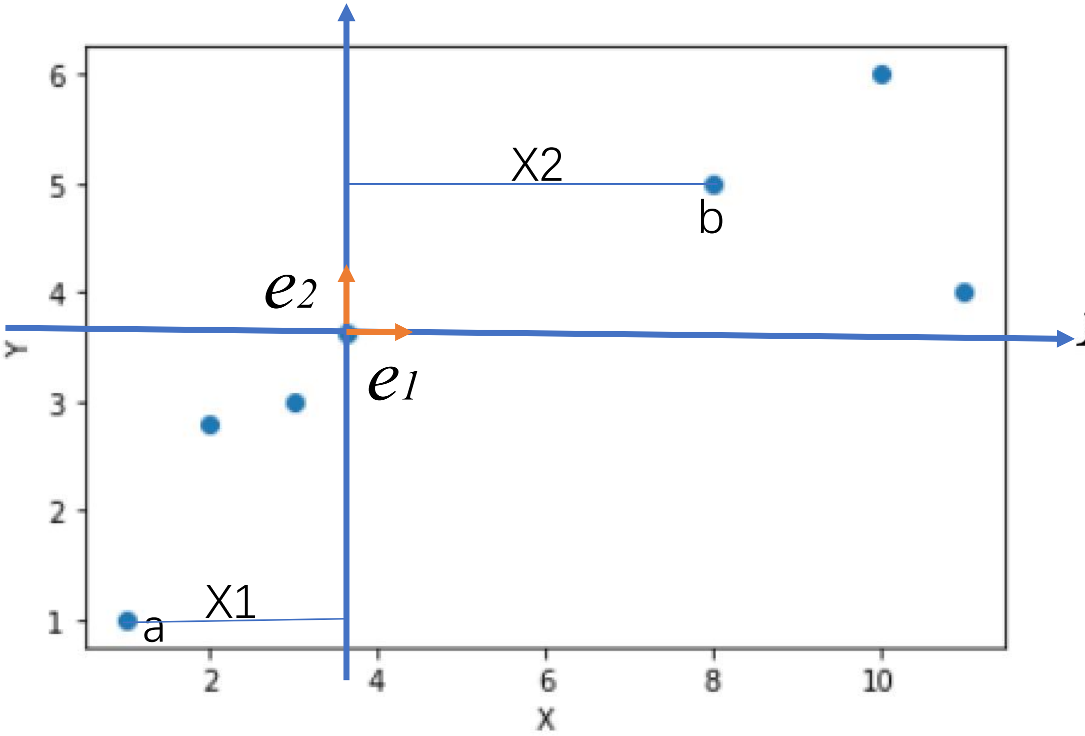
e_1, e_2 is the base vector.

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \qquad \boldsymbol{b} = \begin{pmatrix} a_2 \\ b_2 \end{pmatrix}$$

$$X_1 = \boldsymbol{a} \cdot \boldsymbol{e_1} = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \cdot \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = a_1 e_{11} + b_1 e_{12}$$

e_1

$$X_2 = \boldsymbol{b} \cdot \boldsymbol{e_1} = \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \cdot \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = a_2 e_{11} + b_2 e_{12}$$



$$X_1^2 + X_2^2 = (a_1 e_{11} + b_1 e_{12})^2 + (a_2 e_{11} + b_2 e_{12})^2$$

$$= a_1^2 e_{11}^2 + 2a_1 b_1 e_{11} e_{12} + b_1^2 e_{12}^2 + a_2^2 e_{11}^2 + 2a_2 b_2 e_{11} e_{12} + b_2^2 e_{12}^2$$

$$= (a_1^2 + a_2^2) e_{11}^2 + 2(a_1 b_1 + a_2 b_2) e_{11} e_{12} + (b_1^2 + b_2^2) e_{12}^2$$

maximum $X_1^2 + X_2^2 = \sum_{i=0}^{2} X_i^2$

$$X_1^2 + X_2^2 = \boldsymbol{e_1^{\mathrm{T}}} \underbrace{\begin{pmatrix} a_1^2 + a_2^2 & a_1 b_1 + a_2 b_2 \\ a_1 b_1 + a_2 b_2 & b_1^2 + b_2^2 \end{pmatrix}}_{P} \boldsymbol{e_1} = \boldsymbol{e_1^{\mathrm{T}}} \boldsymbol{P} \boldsymbol{e_1}$$

Second find a maximum value:

$$X_1^2 + X_2^2 = \boldsymbol{e}_\mathbf{1}^\mathrm{T} \underbrace{\begin{pmatrix} a_1^2 + a_2^2 & a_1 b_1 + a_2 b_2 \\ a_1 b_1 + a_2 b_2 & b_1^2 + b_2^2 \end{pmatrix}}_{P} \boldsymbol{e}_\mathbf{1} = \boldsymbol{e}_\mathbf{1}^\mathrm{T} P \boldsymbol{e}_\mathbf{1}$$

symmetric matrix: $P = U \Sigma U^\mathrm{T}$

orthogonal matrix: $UU^\mathrm{T} = I$

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$ I: unit matrix

diagonal matrix: $\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$

σ_1, σ_2 singular value

Second find a maximum value:

$$X_1^2 + X_2^2 = \boldsymbol{e}_1^{\mathrm{T}} P \boldsymbol{e}_1$$

$$= \boldsymbol{e}_1^{\mathrm{T}} U \Sigma U^{\mathrm{T}} \boldsymbol{e}_1$$

$$= (U^{\mathrm{T}} \boldsymbol{e}_1)^{\mathrm{T}} \Sigma (U^{\mathrm{T}} \boldsymbol{e}_1)$$

$$\boldsymbol{n} = U^{\mathrm{T}} \boldsymbol{e}_1 \qquad = \boldsymbol{n}^{\mathrm{T}} \Sigma \boldsymbol{n}$$

$$= \begin{pmatrix} n_1 & n_2 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix}$$

$$= \sigma_1 n_1^2 + \sigma_2 n_2^2$$

$$\boldsymbol{e}_1 = \begin{cases} P = U \Sigma U^{\mathrm{T}} \\ \text{Singular vector corresponding to maximum singular value}(\sigma 1) \end{cases}$$

$$\boldsymbol{e}_2 = \begin{cases} P = U \Sigma U^{\mathrm{T}} \\ \text{Singular vector corresponding to maximum singular value}(\sigma 2) \end{cases}$$

Second find a **maximum value:**

Simplification:

$$X_1^2 + X_2^2 = \sum_{i=0}^{2} X_i^2$$

covariance matrix: $Q = \frac{1}{n-1}\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}$

$$Var(X) = \frac{\sum(X-\overline{X})^2}{n-1}$$

Compare with matrix P:

$$P = \begin{pmatrix} a_1^2 + a_2^2 & a_1 b_1 + a_2 b_2 \\ a_1 b_1 + a_2 b_2 & b_1^2 + b_2^2 \end{pmatrix} = \begin{pmatrix} X \cdot X & X \cdot Y \\ X \cdot Y & Y \cdot Y \end{pmatrix}$$

$$\overline{X} = 0$$

$$Var(X) = \frac{1}{n-1}\sum X_i^2$$

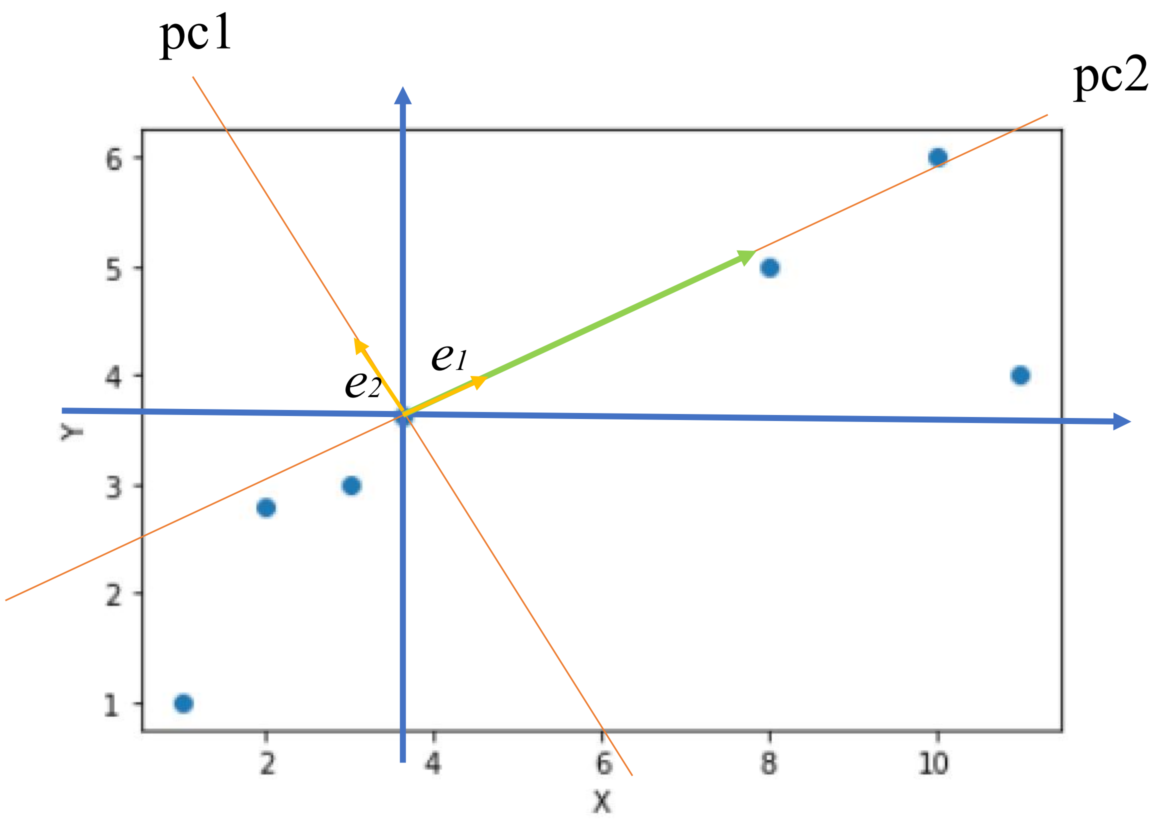
$$Q = \frac{1}{\text{n-1}} P = U \begin{pmatrix} \frac{\sigma_1}{n} & 0 \\ 0 & \frac{\sigma_2}{n} \end{pmatrix} U^{\mathrm{T}}$$

eigenvector

eigenvalue

$$Cov(X,Y) = \frac{1}{\text{n}-1}\sum X_i Y_i$$

Second find a maximum value:



Application in astronomy

To estimate the wavelength-dependent continuum level, we use a principal component analysis (PCA) as described by Eilers et al. (2017). This PCA-based continuum estimate C_λ is used to calculate the Lyα transmitted flux $F_\alpha = e^{-\tau_\alpha}$,

$$F_\alpha = f_\lambda / C_\lambda + n_\lambda / C_\lambda, \tag{3}$$

where f_λ is the observed flux and n_λ is the noise in the Q1148 ESI spectrum.

We find the best fit mean extinction curve and multi-parameter families of extinction curves by finding low-dimensional subspaces of the ten dimensional space of observed reddenings that best explain the data. This procedure is essentially a weighted principal component analysis (PCA), with separate weights (σ^{-2}) for each observation (Jolliffe 2002). We find these low-dimensional subspaces via the Heteroscedastic Matrix Factorization technique of Tsalmantza & Hogg (2012) (see also Gabriel & Zamir 1979; Roweis 1998; Tamuz et al. 2005). This technique, in contrast to classical PCA, appropriately accounts for the heteroscedastic uncertainty in the observations. In analogy with PCA, we call the vectors in these subspaces principal components, and order them according to the first subspace in which they appear.

Useful webpages

Basic knowledgment:

1.Principal Component Analysis (columbia.edu)

2.What Is Principal Component Analysis (PCA) and How It Is Used? (sartorius.com)

3.A Step-by-Step Explanation of Principal Component Analysis (PCA) | Built In

4.https://www.zhihu.com/question/41120789

Python package:

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

That’s all, thank you!