Dimensionality Reduction

Principal Component Analysis

Why reduce dimension?

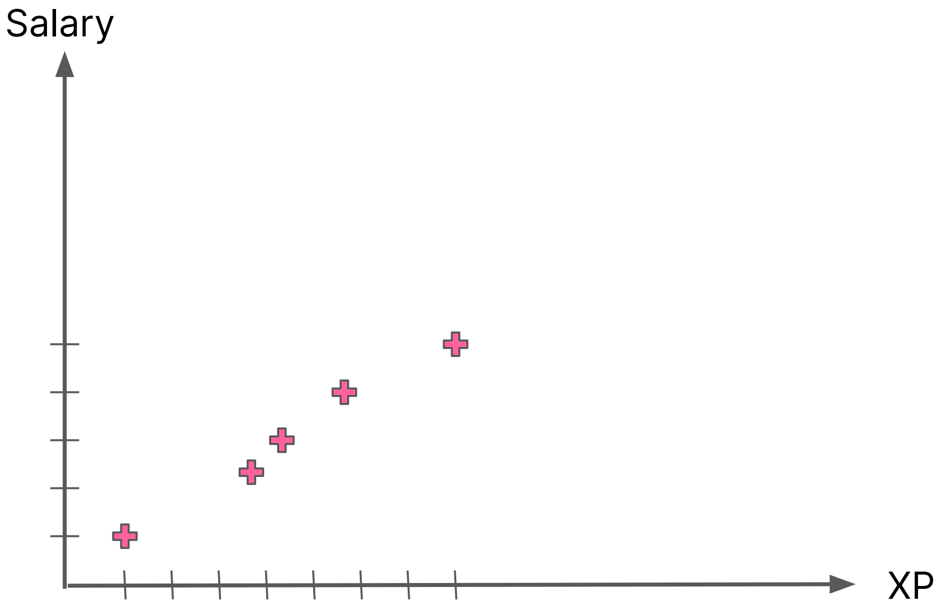
Why use dimensionality reduction algorithms

- **Data visualization** ⇛ Visualize lots of features into a 2-D graph
- **Reduce Noise** ⇛ Reduce redundancy in a data

Data Visualisation

Years of experience	Salary
1	10 000
2	13 000
3	13 500
4	15 000
5	17 000

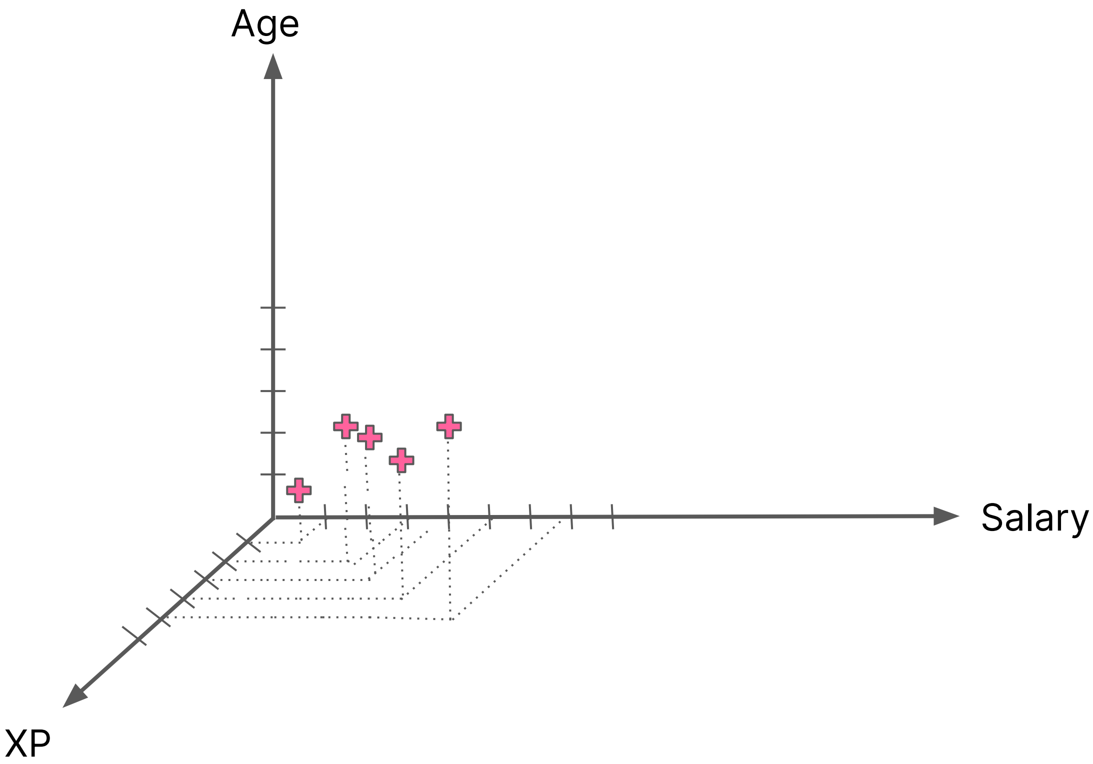
Data Visualisation



Data Visualisation

Years of experience	Salary	Age
1	10 000	25
2	13 000	27
3	13 500	32
4	15 000	29
5	17 000	35

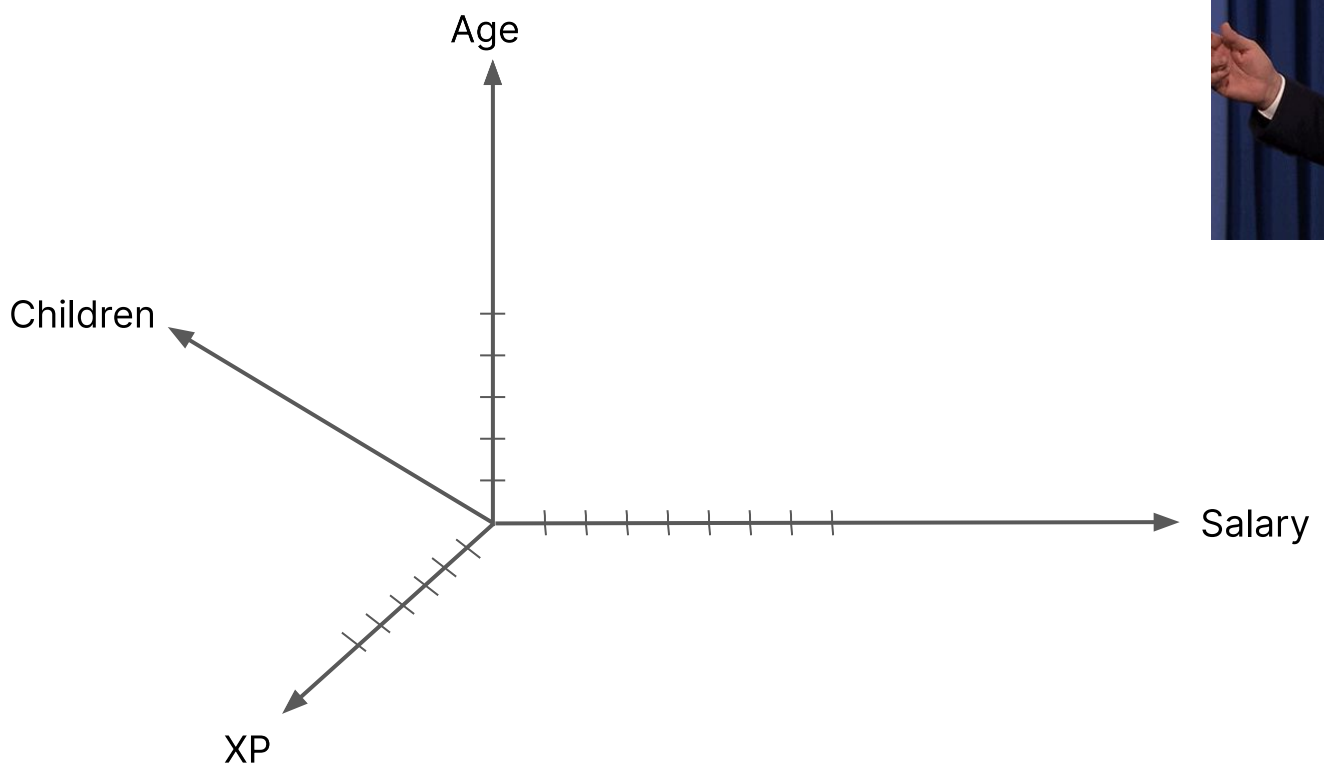
Data Visualisation



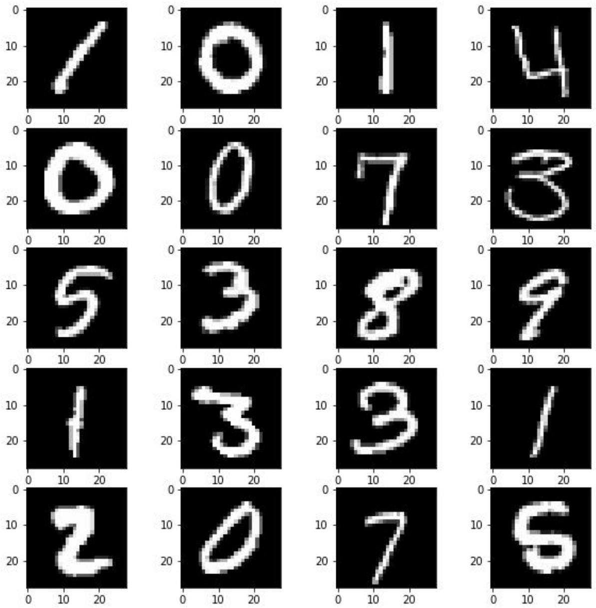
Data Visualisation

Years of experience	Salary	Age	Children
1	10 000	25	0
2	13 000	27	1
3	13 500	32	2
4	15 000	29	0
5	17 000	35	3

Data Visualisation

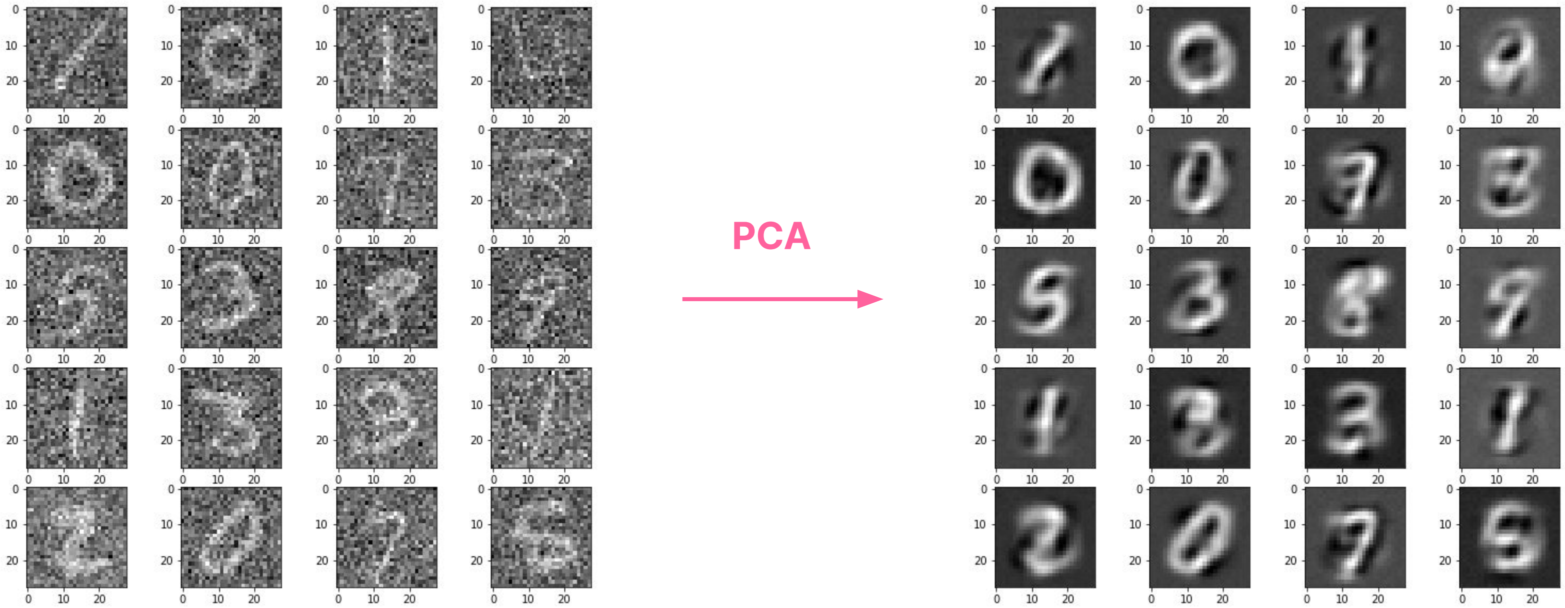


Reduce Noise



Predict
Handwritten digits

Reduce Noise



PCA

Why do PCA?

XP	Salary	Age	Children
1	10 000	25	0
2	13 000	27	1
3	13 500	32	2
4	15 000	29	0
5	17 000	35	3

PC1	PC2
x1	y1
x2	y2
x3	y3
x4	y4
x5	y5

How to do PCA?

XP	Salary	Age	Children
1	10 000	25	0
2	13 000	27	1
3	13 500	32	2
4	15 000	29	0
5	17 000	35	3

PC1	PC2
x1	y1
x2	y2
x3	y3
x4	y4
x5	y5

How to do PCA?

$$\mathrm{x}_1 = a_1 \times \underbrace{1}_{\text{XP}} + a_2 \times \underbrace{10000}_{\text{Salary}} + a_3 \times \underbrace{25}_{\text{Age}} + a_4 \times \underbrace{0}_{\text{Children}}$$

$$\mathrm{x}_2 = a_1 \times 2 + a_2 \times 13000 + a_3 \times 27 + a_4 \times 0$$

$$\mathrm{x}_3 = a_1 \times 3 + a_2 \times 13500 + a_3 \times 32 + a_4 \times 2$$

$$\mathrm{x}_4 = a_1 \times 4 + a_2 \times 15000 + a_3 \times 29 + a_4 \times 0$$

$$\mathrm{x}_5 = a_1 \times 5 + a_2 \times 17000 + a_3 \times 35 + a_4 \times 3$$

How to do PCA?

This is a dot product

$$PC_1 = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 10000 & 13000 & 13500 & 15000 & 17000 \\ 25 & 27 & 32 & 29 & 35 \\ 0 & 1 & 2 & 0 & 3 \end{pmatrix}$$

How to do PCA?

Eigen-vector

$$PCA = \underbrace{\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \end{pmatrix}}_{\text{How do we find these?}} \cdot \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 10000 & 13000 & 13500 & 15000 & 17000 \\ 25 & 27 & 32 & 29 & 35 \\ 0 & 1 & 2 & 0 & 3 \end{pmatrix}$$

Eigen-vector

Process

1. Normalize X
2. Calculate Covariance Matrix
3. Calculate Eigen-vectors & Eigen-values
4. Deduct PCA

Normalize X

Process

1. Normalize X
2. Calculate Covariance Matrix
3. Calculate Eigen-vectors & Eigen-values
4. Deduct PCA

Let's go back to a simple example

Years of experience	Salary
1	10 000
2	13 000
3	13 500
4	15 000
5	17 000

Let’s go back to a simple example

Years of experience	Salary
1	10 000
2	13 000
3	13 500
4	15 000
5	17 000

$$\frac{x_i - \mu}{\sigma}$$

Where:

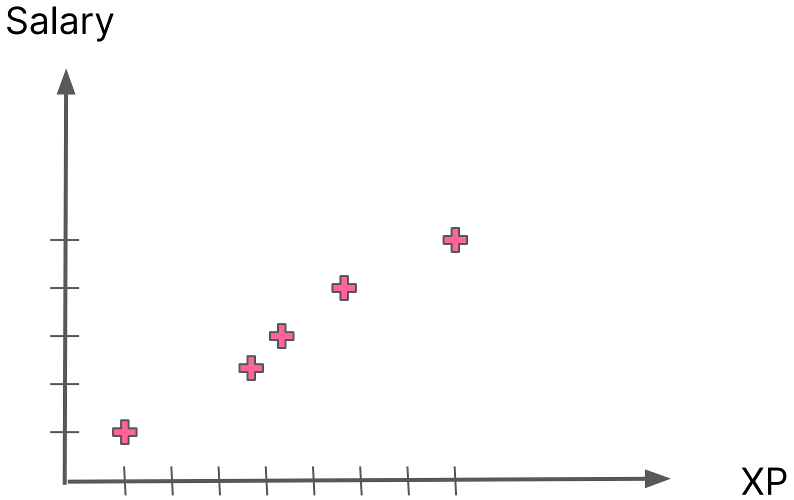
$\mu = mean$

$\sigma = standard\ deviation$

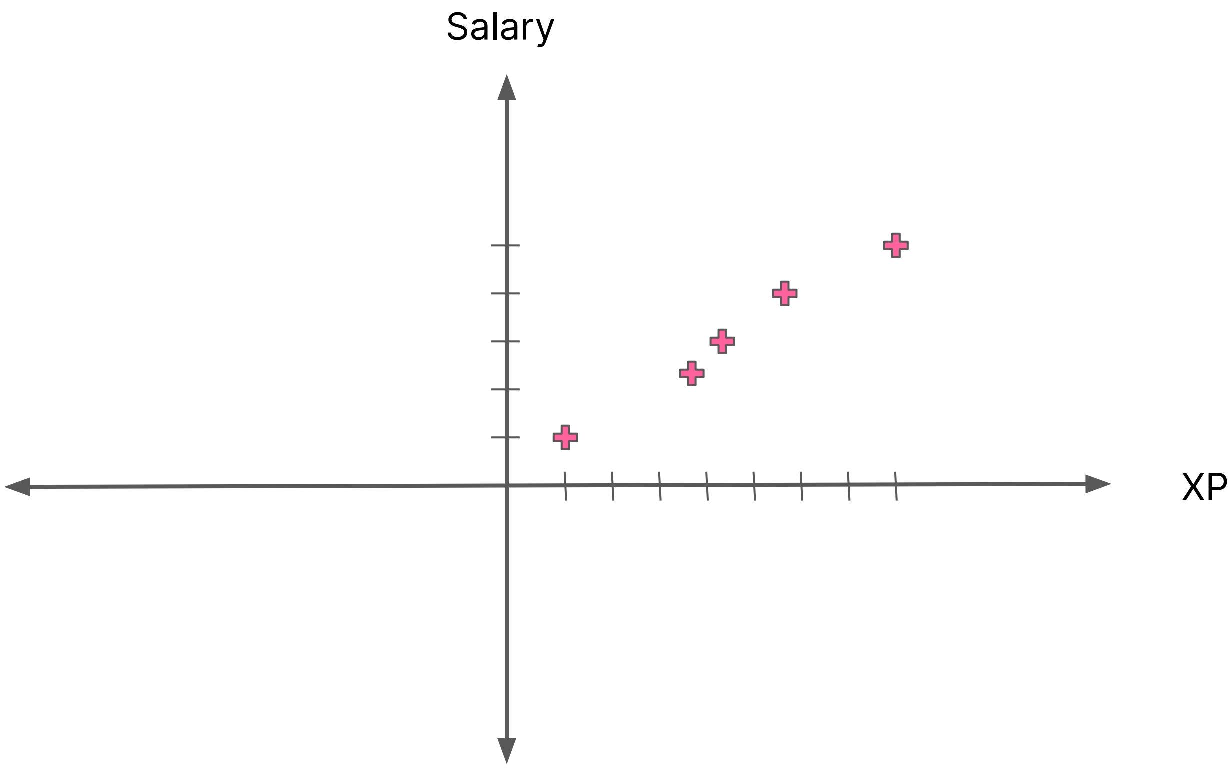
Normalized

Years of experience	Salary
-1.41	-1.60
-0.71	-0.30
0	-0.09
0.71	0.56
1.41	1.42

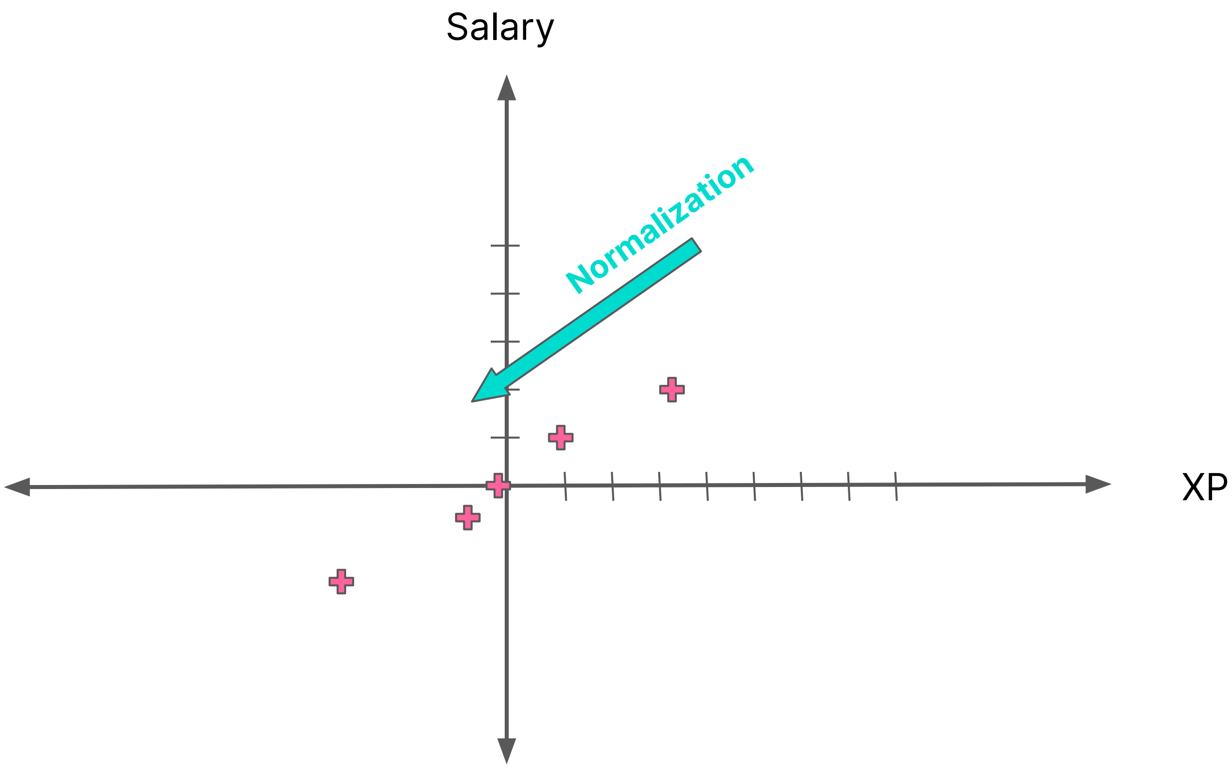
Graphically



Graphically



Graphically



Calculate Covariance Matrix

Process

1. Normalize X
2. Calculate Covariance Matrix
3. Calculate Eigen-vectors & Eigen-values
4. Deduct PCA

Definitions

- **Variance** ⇒ How data points are spread out in a given variable
- **Covariance** ⇒ How two variables are related to each other

Variance & Covariance

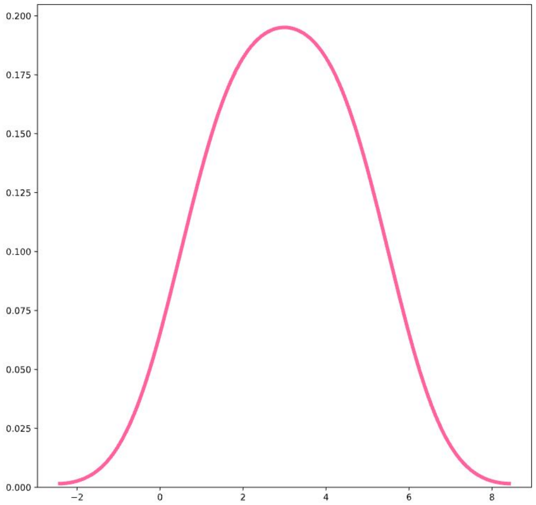
- **Variance** ⇒ How data points are spread out in a given variable

$$\frac{\sum(x_i - \bar{x})^2}{n-1}$$

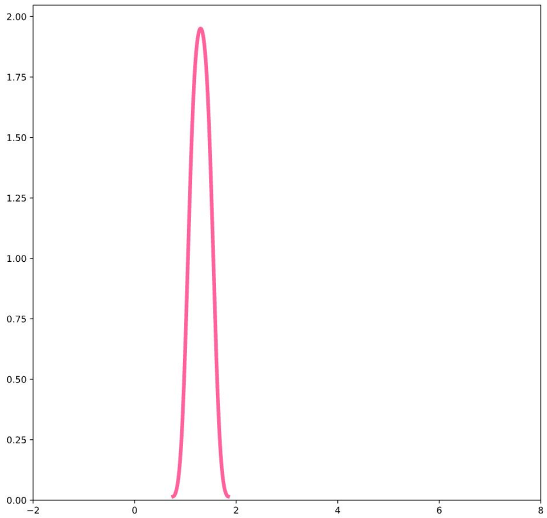
- **Covariance** ⇒ How two variables are related to each other

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Graphically



High Variance



Low Variance

Let's calculate covariance matrix

Variance

Covariance

	XP	Salary
XP	1.25	1.22
Salary	1.22	1.25

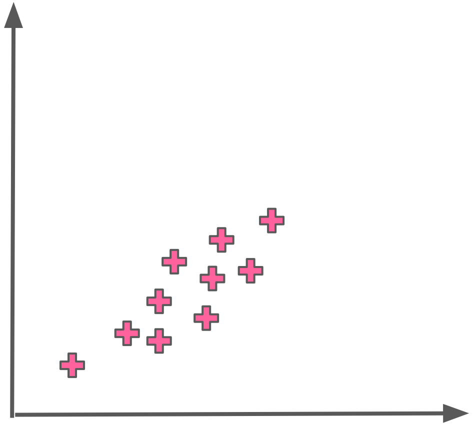
Variance

Covariance

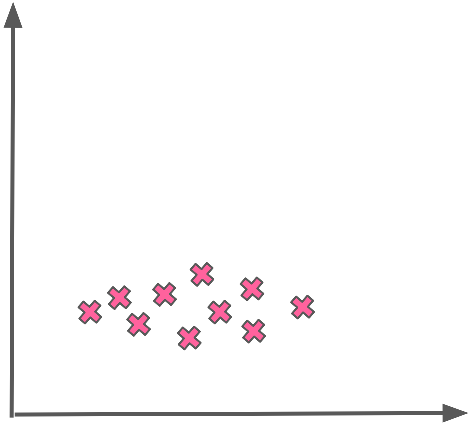
Interpretation

- $cov_{(x,y)} \to \pm\infty$ Statistically **dependent**
- $cov_{(x,y)} \to 0$ Statistically **independent**

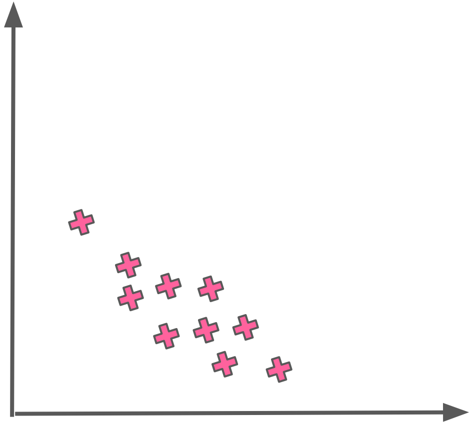
Graphically



Positive Covariance



No Covariance



Negative Covariance

Why using Covariance Matrix?

Process

1. Normalize X
2. Calculate Covariance Matrix
3. Calculate Eigen-vectors & Eigen-values
4. Deduct PCA

Reminder

- We want to **remove redundancy**
- We want to **decompose our dataset into a smaller dataset**

Reminder

Redundancy

- $cov_{(x,y)} \rightarrow \pm\infty$ Statistically **dependent**
- $cov_{(x,y)} \rightarrow 0$ Statistically **independent**

No Redundancy

Let’s calculate covariance matrix

	XP	Salary
XP	1.25	1.22
Salary	1.22	1.25

→

Ideal matrix with no redundancy

	XP	Salary
XP	λ_1	0
Salary	0	λ_2

SVD

Process

1. Normalize X
2. Calculate Covariance Matrix
3. Calculate Eigen-vectors & Eigen-values
4. Deduct PCA

Let’s take our dataset back

Years of experience	Salary
-1.41	-1.60
-0.71	-0.30
0	-0.09
0.71	0.56
1.41	1.42

A

Singular Value Decomposition

$$A = U\Sigma V^{\mathsf{T}}$$

Singular Value Decomposition

$$U = \begin{pmatrix} u_{11} & \cdots & u_{m1} \\ u_{12} & \cdots & u_{m2} \\ \vdots & \ddots & \vdots \\ u_{1m} & \cdots & u_{mm} \end{pmatrix}$$

Eigen Vectors of AA^{T}

$$\Sigma = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & & 0 \\ 0 & \sqrt{\lambda_2} & & & \vdots \\ \vdots & & \ddots & & \\ & & & \sqrt{\lambda_r} & \\ & & & & \ddots \\ 0 & \cdots & & & 0 \end{pmatrix}$$

Eigen Values of AA^{T} and $A^{\mathsf{T}}A$

$$V^{\mathsf{T}} = \begin{pmatrix} v_{11} & \cdots & v_{n1} \\ v_{12} & \cdots & v_{n2} \\ \vdots & \ddots & \vdots \\ v_{1n} & \cdots & v_{nn} \end{pmatrix}$$

Eigen Vectors of $A^{\mathsf{T}}A$

Let’s take our example

Years of experience	Salary
-1.41	-1.60
-0.71	-0.30
0	-0.09
0.71	0.56
1.41	1.42

Let’s take our example

XP	-1.41	-0.71	0	0.71	1.41
Salary	-1.60	-0.30	-0.09	0.56	1.42

A

Let's take our example

$$AA^{\mathsf{T}} = \begin{pmatrix} 1.25 & 1.22 \\ 1.22 & 1.25 \end{pmatrix}$$

Covariance Matrix!!

$$A^{\mathsf{T}}A = \begin{pmatrix} 4.55 & 1.48 & 0.14 & -1.90 & -4.27 \\ 1.48 & 0.59 & 0.26 & -0.67 & -1.43 \\ 0.14 & 0.26 & 0.007 & -0.048 & -0.12 \\ -1.90 & -0.67 & -0.05 & 0.81 & 1.80 \\ -4.27 & -1.43 & -0.12 & 1.80 & 4.03 \end{pmatrix}$$

How to find Eigen Vectors & Eigen Values?

$$AX = \lambda X$$

$$AX - \lambda X = 0$$

$$\boxed{(A - \lambda)X = 0}$$

Linear Algebra problem - **can be solved using Numpy**

This is what we can use to find lambda

Let’s find Eigen values

$$AA^{\mathsf{T}}X = \lambda X$$

$$AA^{\mathsf{T}}X - \lambda X = 0$$

$$\boxed{(AA^{\mathsf{T}} - \lambda)X = 0}$$

Let's find Eigen values

$$(AA^{\mathsf{T}} - \lambda) = \begin{pmatrix} 1.25 & 1.22 \\ 1.22 & 1.25 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

$$(AA^{\mathsf{T}} - \lambda) = \begin{pmatrix} 1.25 - \lambda & 1.22 \\ 1.22 & 1.25 - \lambda \end{pmatrix}$$

Should be equal to 0

Let's find Eigen values

$$(AA^{\mathsf{T}} - \lambda) = \begin{pmatrix} 1.25 - \lambda & 1.22 \\ 1.22 & 1.25 - \lambda \end{pmatrix}$$

$$det(AA^{\mathsf{T}} - \lambda) = (1.25 - \lambda)^2 - 1.22^2$$

$$\boxed{det(AA^{\mathsf{T}} - \lambda) = 0.07 - 2.5\lambda + \lambda^2}$$

Let's find Eigen values

$$\lambda_1 = 9.89$$

$$\lambda_2 = 0.11$$

$$\sqrt{\lambda_1} = 3.14$$

$$\sqrt{\lambda_2} = 0.34$$

Let’s find Eigen Vectors

$$AA^{\top}\underbrace{X} = \lambda\underbrace{X}$$

Eigen Vectors

Let’s find Eigen Vectors

$$(AA^{\mathsf{T}} - \lambda)X = 0$$

$$(AA^{\mathsf{T}} - \lambda)X = (\begin{pmatrix} 1.25 & 1.22 \\ 1.22 & 1.25 \end{pmatrix} - \begin{pmatrix} 0.11 & 0 \\ 0 & 9.89 \end{pmatrix}) \cdot \begin{pmatrix} x \\ y \end{pmatrix}$$

Let's find Eigen Vectors

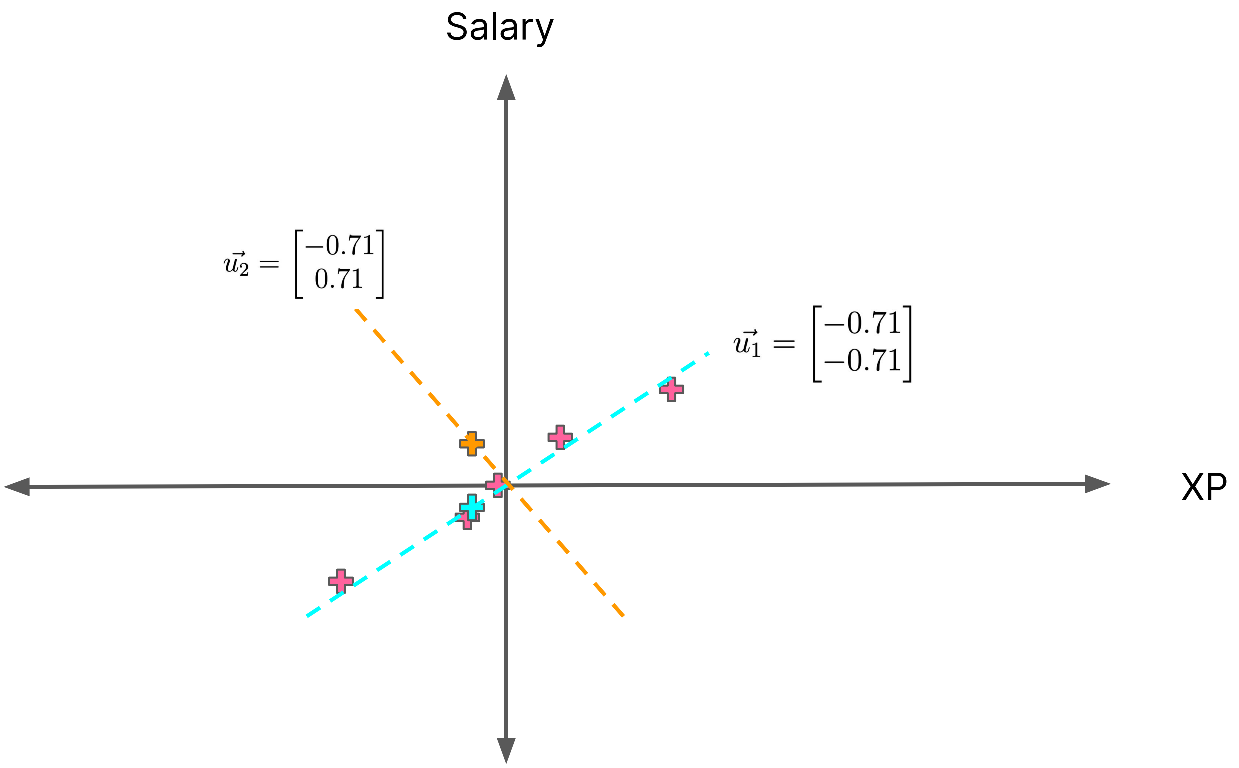
$$1.14x + 1.22y = 0$$

$$1.22x + -8.64y = 0$$

Fast Forward

$$U = \underbrace{\begin{pmatrix} -0.71 & -0.71 \\ -0.71 & 0.71 \end{pmatrix}}_{\text{Eigen-vectors}} \quad \Sigma = \underbrace{\begin{pmatrix} 3.14 & 0 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 & 0 \end{pmatrix}}_{\text{Eigen-values}} \quad V^{\top} = \begin{pmatrix} -0.68 & 0.39 & -0.07 & 0.17 & 0.60 \\ -0.23 & -0.85 & 0.17 & 0.39 & 0.20 \\ -0.02 & 0.18 & 0.98 & -0.03 & -0.00 \\ 0.28 & 0.31 & -0.02 & 0.90 & -0.13 \\ 0.64 & -0.02 & 0.02 & -0.08 & 0.76 \end{pmatrix}$$

Graphically



Deduct PCA

Process

1. Normalize X
2. Calculate Covariance Matrix
3. Calculate Eigen-vectors & Eigen-values
4. Deduct PCA

Reminder

$$U = \underbrace{\begin{pmatrix} -0.71 & -0.71 \\ -0.71 & 0.71 \end{pmatrix}}_{\text{Eigen-vectors}} \quad \Sigma = \underbrace{\begin{pmatrix} 3.14 & 0 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 & 0 \end{pmatrix}}_{\text{Eigen-values}} \quad V^\top = \begin{pmatrix} -0.68 & 0.39 & -0.07 & 0.17 & 0.60 \\ -0.23 & -0.85 & 0.17 & 0.39 & 0.20 \\ -0.02 & 0.18 & 0.98 & -0.03 & -0.00 \\ 0.28 & 0.31 & -0.02 & 0.90 & -0.13 \\ 0.64 & -0.02 & 0.02 & -0.08 & 0.76 \end{pmatrix}$$

Reminder

$$\underbrace{U = \begin{pmatrix} -0.71 & -0.71 \\ -0.71 & 0.71 \end{pmatrix}}_{\text{Principal Components}} \quad \underbrace{\Sigma = \begin{pmatrix} 3.14 & 0 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 & 0 \end{pmatrix}}_{\text{Eigen-values}} \quad V^\top = \begin{pmatrix} -0.68 & 0.39 & -0.07 & 0.17 & 0.60 \\ -0.23 & -0.85 & 0.17 & 0.39 & 0.20 \\ -0.02 & 0.18 & 0.98 & -0.03 & -0.00 \\ 0.28 & 0.31 & -0.02 & 0.90 & -0.13 \\ 0.64 & -0.02 & 0.02 & -0.08 & 0.76 \end{pmatrix}$$

Let's compute Principal Components

$$AU$$

Let's compute Principal Components

$$AU = \begin{pmatrix} 2.13 & -0.13 \\ 0.71 & 0.29 \\ 0.06 & -0.06 \\ -0.90 & -0.10 \\ -2.00 & 0.01 \end{pmatrix}$$

Let's check our new covariance

$$cov(AU)$$

	PC1	PC2
PC1	2.47	0
PC2	0	0.03

No Redundancies!!

How to reduce dimension?

	PC1	PC2
PC1	2.47	0
PC2	0	0.03

Variance explained

	PC1	PC2
PC1	99%	0
PC2	0	0.1%

Variance explained ratio

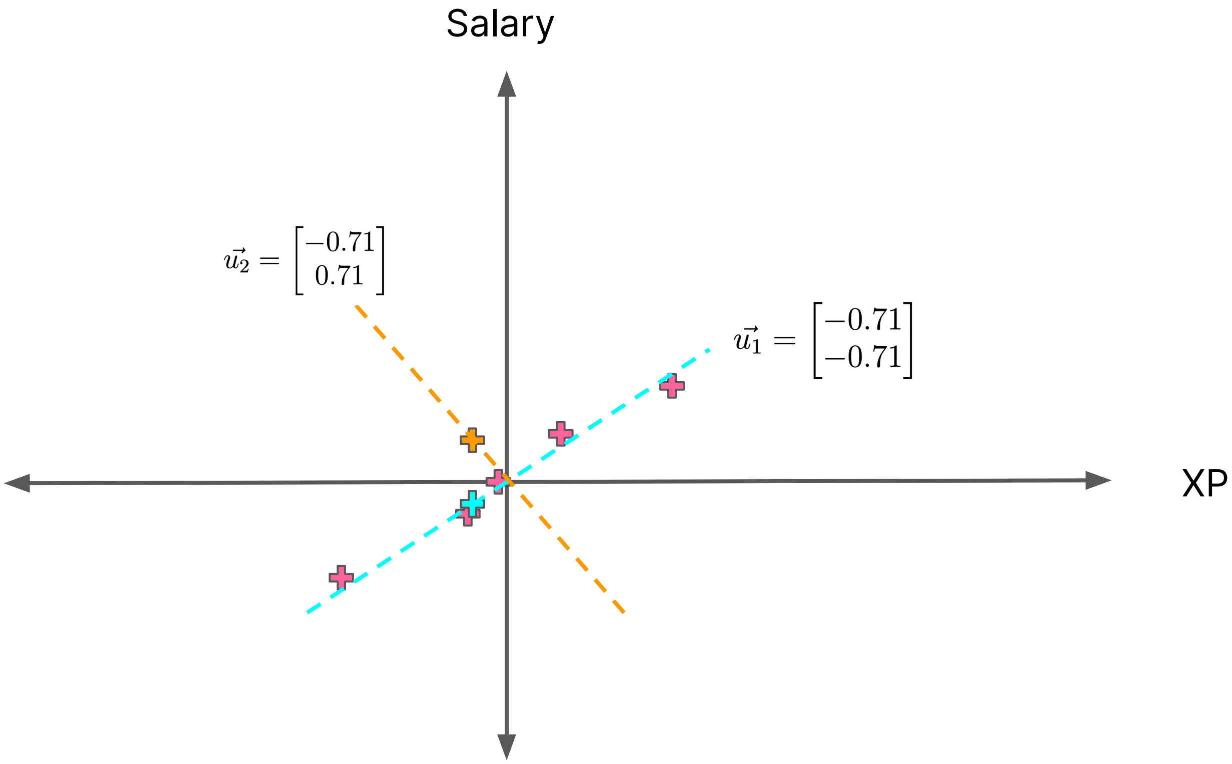
How to reduce dimension?

	PC1	PC2
PC1	2.47	0
PC2	0	0.03

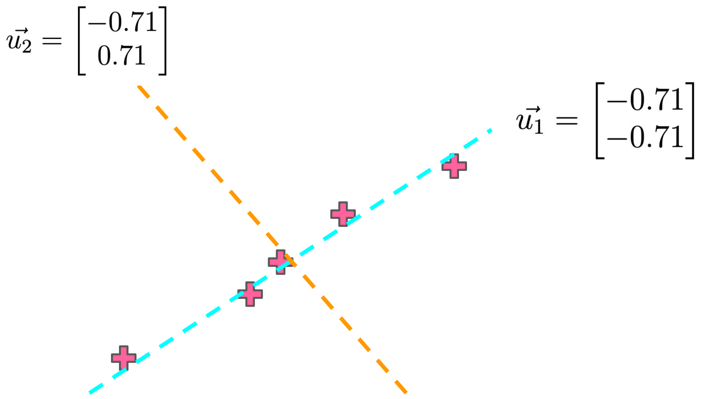
	PC1	PC2
PC1	99%	0
PC2	0	0.1%

We can keep only PC1 !

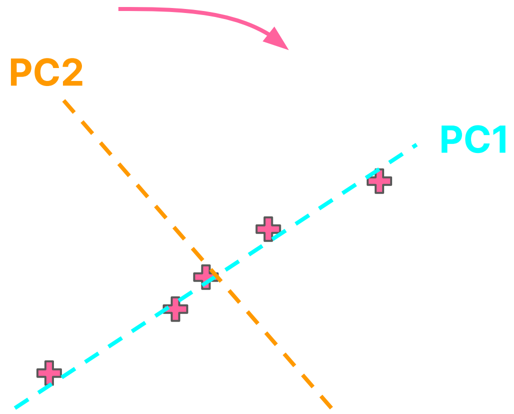
Graphically



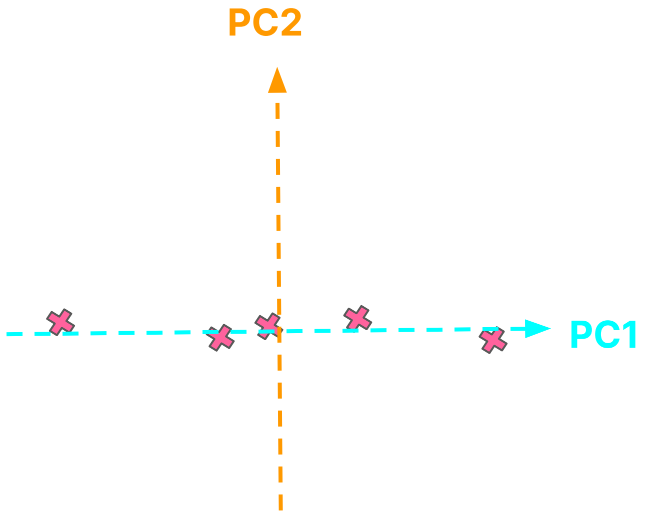
Graphically



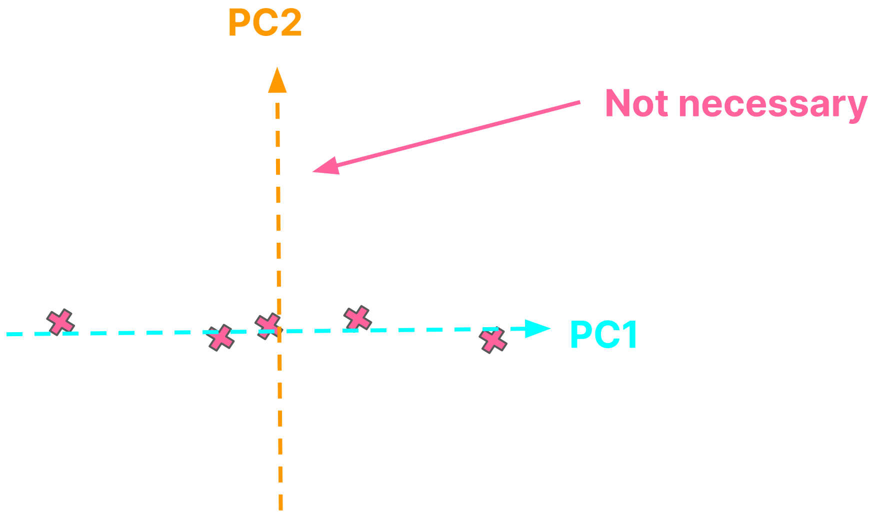
Graphically



Graphically



Graphically



Graphically

Thanks!

See you in the next course