Principal Component Analysis

From riches to rags

PCA Example

PCA – How it works

Step 1: Standardization of features (independent variables only)

Step 2: Calculate Covariance Matrix

Step 3: Compute Eigen Values & Eigen Vectors of Covariance Matrix
Then Assign the Principal Components Accordingly

Step 4: Keep Important PCs and discard less significant ones

Step 5: Create Final Dataset

PCA – How it works

Suppose below is our data:

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

We now wish to do a PCA on this one and see if we can reduce it to lower dimension.

Note: Here for the sake of simplicity we took 2 dimensions (features). However if there are more dimensions, then we follow the same approach that we will see now.

PCA – How it works

Step 1: Standardize your data.

Make sure you do this step as it is a good practice. However if your data are already in good scales then you may skip it.

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Here, our data is already in good scale (X and Y) hence we skip this step.

PCA – How it works

Step 2: Calculate co-variance matrix
The covariance matrix for our data will be like

C =

cov(x, x)	cov(x, y)
cov(y, x)	cov(y, y)

Note: We have only 2D data, hence the covariance matrix is 2x2. If we have N-D data, then our covariance matrix will be nxn

e.g. Covariance for 3D data will be like:

cov(x,x)	cov(x,y)	cov(x,z)
cov(y,x)	cov(y,y)	cov(y,z)
cov(z,x)	cov(z,y)	cov(z,z)

PCA – How it works

Step 2: Calculate co-variance matrix

The covariance matrix for our data will be like

C =

cov(x, x)	cov(x, y)
cov(y, x)	cov(y, y)

Covariance Formula is:

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

PCA – How it works

Step 2: Calculate co-variance matrix

The covariance matrix for our data will be like

C =

cov(x, x)	cov(x, y)
cov(y, x)	cov(y, y)

After applying the formula,
we get our covariance matrix

0.6165	0.6154
0.6154	0.7165

PCA – How it works

Step 3: Compute Eigen Values & Eigen Vectors

To get the **Eigen values**, we have to solve for the determinants of :

C – λI = 0

Where,

C = Covariance Matrix

λ = Constant/Eigen value

I = Identity Matrix

Thus

$$\begin{vmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 0$$

$$\begin{vmatrix} 0.6165 - \lambda & 0.6154 \\ 0.6154 & 0.7165 - \lambda \end{vmatrix} = 0$$

PCA – How it works

Step 3: Compute Eigen Values & Eigen Vectors

Eventually we get the equation:
$(0.6165-\lambda) * (0.7165-\lambda) – 0.6154 * 0.6154 = 0$

We will get a quadratic equation here,
Which means it has 2 roots.

After solving, we get the roots as:
$\boldsymbol{\lambda_1 = 0.4908}$ **&** $\boldsymbol{\lambda_2 = 1.2480}$

Hence our eigen values are:
$\lambda_1 = 0.4908$ & $\lambda_2 = 1.2480$

PCA – How it works

Step 3: Compute Eigen Values & Eigen Vectors
To get the **Eigen vectors**, we have to solve for the determinants of :

A X = 0

$$\begin{bmatrix} 0.6165 - \lambda & 0.6154 \\ 0.6154 & 0.7165 - \lambda \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \end{bmatrix} = 0$$

We need to solve for both Eigen Values :
$\lambda_1 = 0.4908$ & $\lambda_2 = 1.2480$

PCA – How it works

Step 3: Compute Eigen Values & Eigen Vectors

Solving for:
$\lambda_1 = 0.4908$

0.6165 – 0.4908	0.6154	X_1	= 0
0.6154	0.7165 – 0.4908	Y_1	

Hence, we get

$0.1257\, X_1 + 0.6154\, Y_1 = 0$
$0.6154\, X_1 + 0.2257\, Y_1 = 0$
Solving above equation we get Eigen vector v_1 for $\lambda_1 = 0.4908$

v_1 =

- 0.735
0.677

PCA – How it works

Step 3: Compute Eigen Values & Eigen Vectors

Solving for:
$\lambda_2 = 1.2840$

$$\begin{bmatrix} 0.6165 - 1.2840 & 0.6154 \\ 0.6154 & 0.7165 - 1.2840 \end{bmatrix} \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} = 0$$

Solving above equation:
we get Eigen vector v_2 for $\lambda_2 = 1.2840$

$$v_2 = \begin{bmatrix} -6.77 \\ -0.735 \end{bmatrix}$$

PCA – How it works

Step 3: Compute Eigen Values & Eigen Vectors

Assigning Principal Components

Here,
λ_2 (1.2480) < λ_1 (0.4908)

Hence,
λ_2 is our Principal Component 1 (PC1)
λ_1 is our Principal Component 2 (PC2)

Corresponding values will be our Eigen vectors:

v1	v2
- 0.735	-0.677
0.677	-0.735

PCA – How it works

Step 4: Keep Important Components

Here,

We keep all the components

Step 5: Create final dataset

Final Dataset = $(\text{FeatureVector})^T$ – $(\text{Standardize Data})^T$

where

T = Transpose

Feature Vector = Vector(Dataframe) you get having the PCs' in it

(PC1 & PC2 here)