Principal Component Analysis

Agenda

- What is PCA?
- Steps Involved in PCA
- Applications of PCA
- Mathematical Illustration
- Coding PCA in python

What is PCA?

- It is a dimensionality-reduction method
  - It reduces the larger data sets into smaller one by reducing the number of variables
  - Still retains most of the information.
  - Trade off for accuracy, as the no. of variables are reduced.
  - Simple to use for Machine Learning algorithms

Steps involved in PCA

- Standardization
- Covariance Matrix Computation
- Computation Of The Eigenvectors And Eigenvalues Of The Covariance Matrix
- Identification Of Principal Components
- Feature Vector
- Recast the data along the principal components axes

Standardization

- Standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

- Why Standardization?
  - To avoid biased results
  - large differences between the ranges of initial variables will dominate over those with small ranges
  - Example - a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1
  - To avoid this transform the data to comparable scales

Cont..

- Tansform the data to comparable scales using the following equation

$$Z = \frac{Value - Mean}{Standard\ Deviation}$$

- After transformation all the variables will be in the same scale

Covariance Matrix Computation

- It gives the understanding of how the variables in the input data set are varying from the mean, with respect to each other.
- Identifies the Highly redundant information – Highly correlated
- To identify correlations – Compute Co variance Matrix
- covariance of a variable with itself is nothing but variance
  - (Cov(a,a)=Var(a))

Covariance Matrix Computation (cont..)

- What do we understand from covariance?
- Significance lies in the Sign.
  - +ve the two variables increase or decrease together (correlated)
  - - Ve One increases when the other decreases (Inversely correlated)
- Covariance Matrix correlations between all the possible pairs of variables

Covariance Matrix Computation (cont..)

- Covariance Matrix is symmetric with dimensions pxp
  - Where p is the number of variables.
- Consider a 3-Dimensional data set with vaiables (x,y,z).
- Covariance Matrix

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

$cov_{x,y}$ = covariance between variable a and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values

Covariance Matrix Computation (cont..)

Covariance Matrix

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

- covariance satisfies commutative property (Cov(a,b)=Cov(b,a)),
- Therefore with respect to main diagonal the upper and lower triangular portions are equal

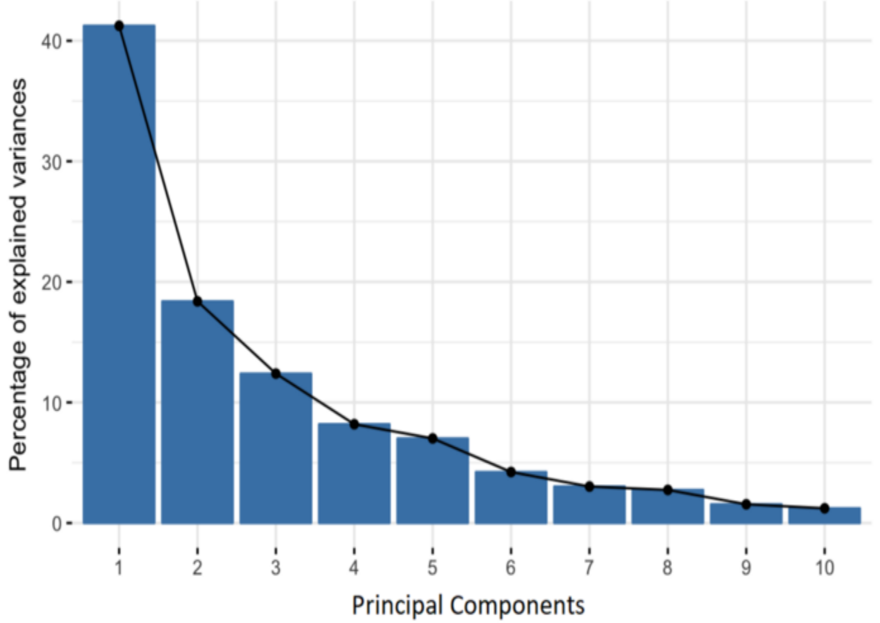
$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Computation Of The Eigenvectors And Eigenvalues Of The Covariance Matrix

- Eigen vector is the projected vector from the data which is perpendicular to the data.
- They are generally projected in the direction of most significant data.
- Eigen vectors are generally column vectors
- Eigen values and Eigen vectors are computed from covariance matrix to determine the principal components.
- the eigenvectors are ordered by their eigen values in descending order, -- this helps in find the principal components in the order of significance.
- The Eigen vector with the highest eigen value is the principle component

Computation Of The Eigenvectors And Eigenvalues Of The Covariance Matrix

- Constructed in such a way that the resultant is uncorrelated
- Most of the information within the initial variables is compressed into the first components.
- If a 10 dimensional data produces 10 principal components then PCA gives maximum information in the first component and remaining information in the other components and so on..



- First principal component -- **largest possible variance** in the data set

- Second principal component -- next highest variance(uncorrelated with the first principal component)

Feature vector

- From the ordered Eigen vectors - choose whether to keep all these components or discard those of lesser significance (of low eigenvalues)
- The remaining matrix of Eigen vectors is called Feature Vector.
- Feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep.
- if we choose to keep only $\boldsymbol{p}$ eigenvectors out of $\boldsymbol{n}$, the final data set will have only $\boldsymbol{p}$ dimensions

RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES

- Use the feature vectors, to reorient the data from the original axes to the ones represented by the principal components (hence the name Principal Components Analysis).

- This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

Cont..

- row feature vector - data
- Row data adjust: - eigen vector
- Multiply the row feature vector into the row data adjust.
- In the result each and every data is converted into the principle component.
- Getting the original data back:

Row original data = (Row feature vector ^{T} x Final Data)+ original Mean

Applications of PCA

- **Data Visualization**
- **Speeding Machine Learning (ML) Algorithm**

Mathematical illustration

Consider the following data set

x	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
y	2.4	0.7	2.9	2.2	3	2.7	1..6	1.1	1.6	0.9

Shifting the data points towards origin by subtracting from mean

x	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
y	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01

Mathematical illustration

Covariance Matrix = [0.61655556 0.61544444]
[0.61544444 0.71655556]]

Since the non diagonal elements in this covariance matrix are positive , x and Y variables will increase together

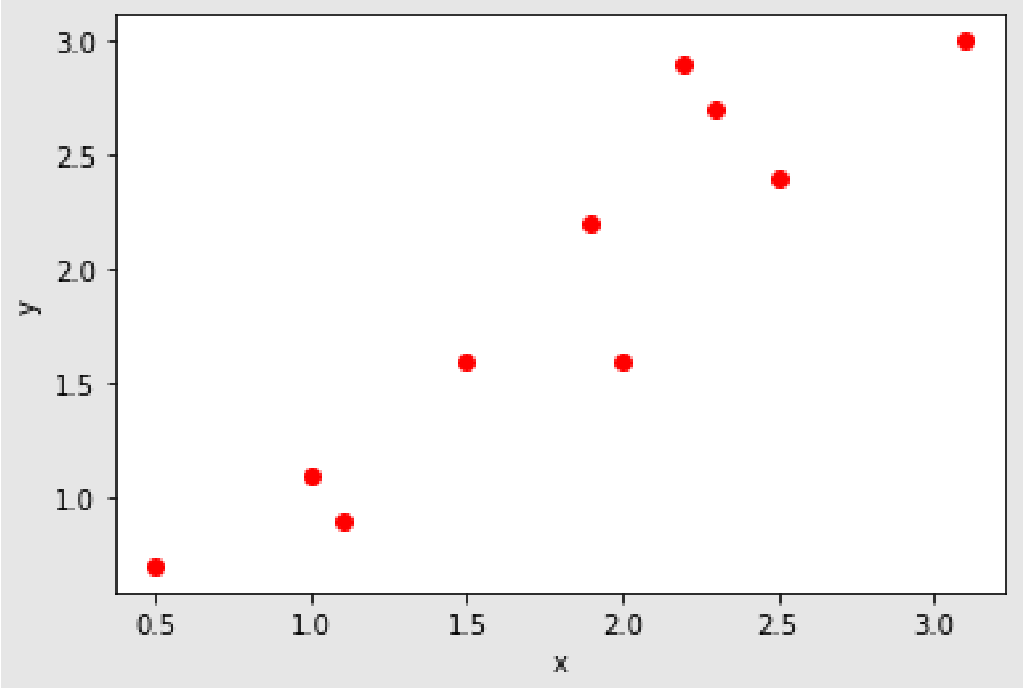
Calculation of Eigen values and Eigen vectors

Eigen Values **[0.0490834 1.28402771]**

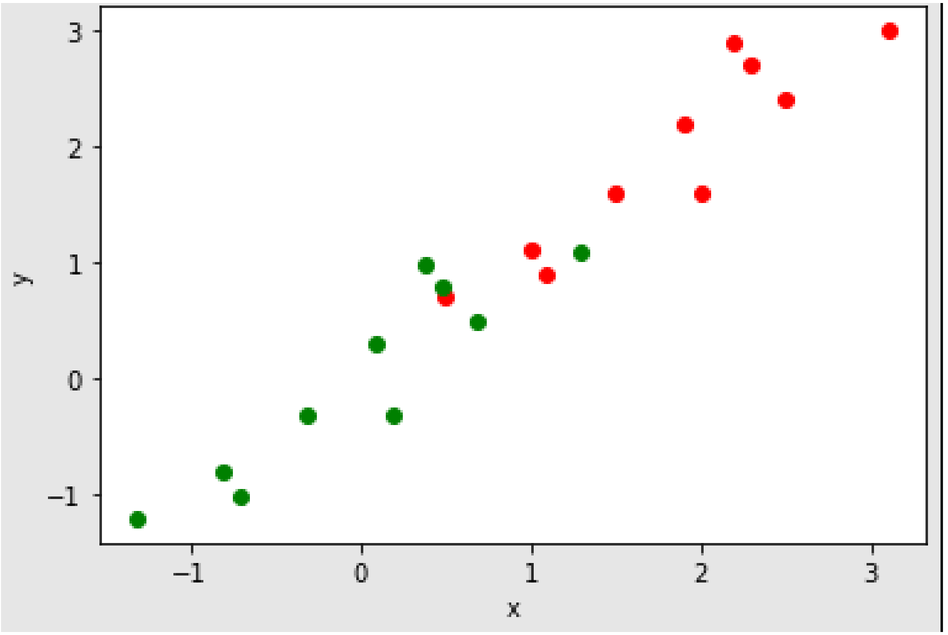
$$\text{eigenvectors} = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Mathematical illustration

Raw data set



Data set Translated to origin



Mathematical illustration

$$\text{eigenvalues} = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Sorted Eigen Values and Eigen Vectors

$$\begin{pmatrix} 1.28402771 \\ .0490833989 \end{pmatrix} \quad \begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

Feature Vector

n=1

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

n=2

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

Mathematical illustration

Final data

For n=1

[0.82797019]
[-1.77758033]
[0.99219749]
[0.27421042]
[1.67580142]
[0.9129491]
[-0.09910944]
[-1.14457216]
[-0.43804614]
[-1.22382056]

n=2
[0.82797019 -0.17511531]
[-1.77758033 0.14285723]
[0.99219749 0.38437499]
[0.27421042 0.13041721]
[1.67580142 -0.20949846]
[0.9129491 0.17528244]
[-0.09910944 -0.3498247]
[-1.14457216 0.04641726]
[-0.43804614 0.01776463]

Final data