# PRINCIPAL COMPONENT ANALYSIS

# Introduction

- PCA aims to reduce the number of variables of a data set, while preserving as much information as possible.

- Principle Component Analysis is closely related to Singular Value Decomposition.

- PCA finds a linear projection of high dimensional data into a lower dimensional subspace such as:
  - The variance retained is maximized.
  - The least square reconstruction error is minimized.

# Steps in performing PCA:

1. Standardization of variables.

Let X1, X2, …., Xp be the variables under study.  p = number of variables.

In case the variables are having different scales, standardize them.

- The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
- If there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges.
- For example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1, which will lead to biased results.
- So, transforming the data to comparable scales can prevent this problem.

## Steps in performing PCA:

1. **Standardization of variables.**

**Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.**

$$z = \frac{X - \mu}{\sigma}$$

**Once the standardization is done, all the variables will be transformed to the same scale.**

# Steps in performing PCA:

## 2. Covariance Matrix Computation

- The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them.
- Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix.
- The covariance matrix is a p × p symmetric matrix (where p is the number of dimensions) that has as entries the covariance associated with all possible pairs of the initial variables.
- For example, for a 3-dimensional data set with 3 variables x, y, and z, the covariance matrix is a 3×3 matrix of this from:

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

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

## Steps in performing PCA:

### 3. Compute The Eigenvectors And Eigenvalues Of The Covariance Matrix To Identify The Principal Components

- **Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data. Before getting to the explanation of these concepts, let’s first understand what do we mean by principal components.**

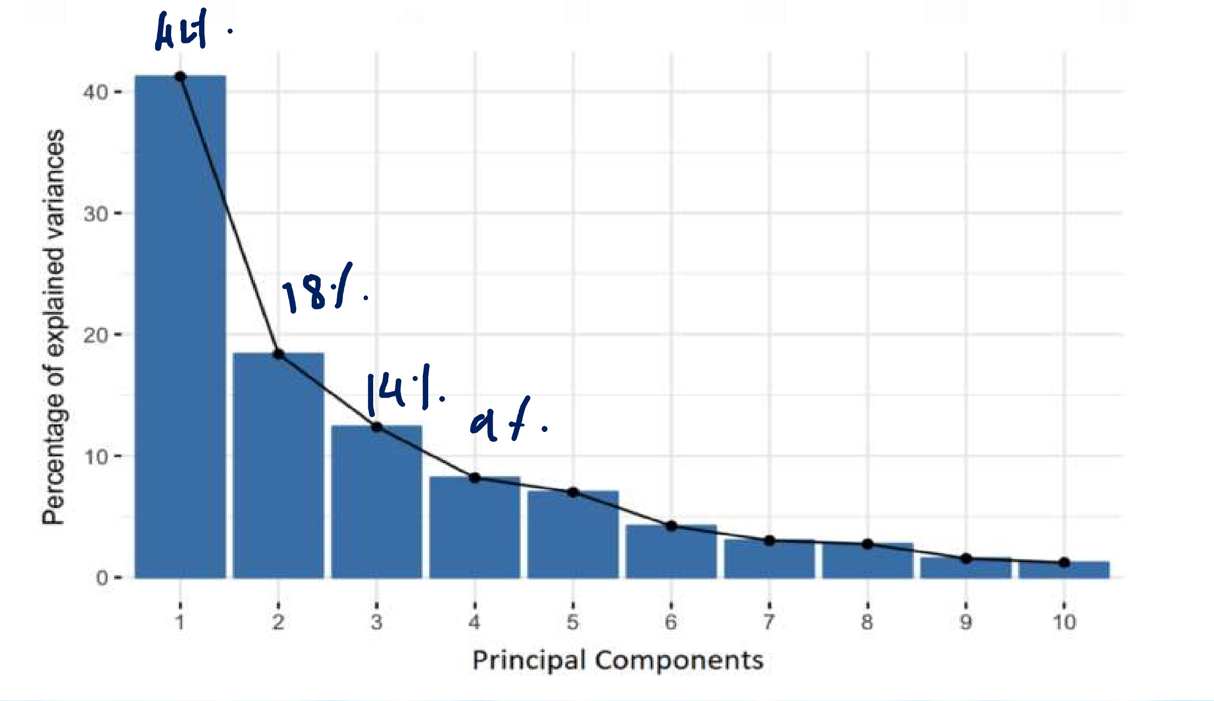
## Steps in performing PCA:

### 3. Compute The Eigenvectors And Eigenvalues Of The Covariance Matrix To Identify The Principal Components

- Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.

# Steps in performing PCA:

## 3. Compute The Eigenvectors And Eigenvalues Of The Covariance Matrix To Identify The Principal Components



$$42 + 18 + 14 = 74\% + 9\% = \underbrace{83\%}$$

## Principal Component Analysis

- Principal components represent the directions of the data that explain a maximal amount of variance, that is to PCs are the lines that capture most information of the data.

- The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has.

- There are as many principal components as there are many variables in the data. So, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set.

## Principal Component Analysis

- The first principal component is the line in which the projection of the points is the most spread out. Or mathematically speaking, it’s the line that maximizes the variance (the average of the squared distances from the projected points to the origin).
- The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.
- This continues until a total of p principal components have been calculated, equal to the original number of variables.

# Principal Component Analysis

- Let $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p$ be the eigenvalues and $e_1, e_2, \cdots e_p$ e the corresponding eigenvectors.

- Let $e_i = (e_1, e_2, \cdots e_p)^T$

- Calculation of Principal components:

$$Y_1 = e_1' X = e_{11}X_1 + e_{21}X_2 + ........ + e_{p1}X_p$$

$$Y_2 = e_2' X = e_{12}X_1 + e_{22}X_2 + ........ + e_{p2}X_p$$

$$\cdot$$

$$\cdot$$

$$\cdot$$

$$Y_p = e_p' X = e_{1p}X_1 + e_{2p}X_2 + ........ + e_{pp}X_p$$

# Principal Component Analysis

- This means principal components the new variables created which are expressed as linear combination of the original variables in the data.
- It can be shown that

$$Var(Yi) = \lambda_i, \ \mathrm{cov}(Y_i, Y_j) = 0 \ \ for \ i \neq j$$

$$corr(Y_i, X_k) = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \ , \ where \ \ \sigma_{kk} = \mathrm{var}(X_i)$$

variance $(Y_i)$ = Eigen values

$Y_1 \neq Y_2$

$Y_2 \neq Y_3$

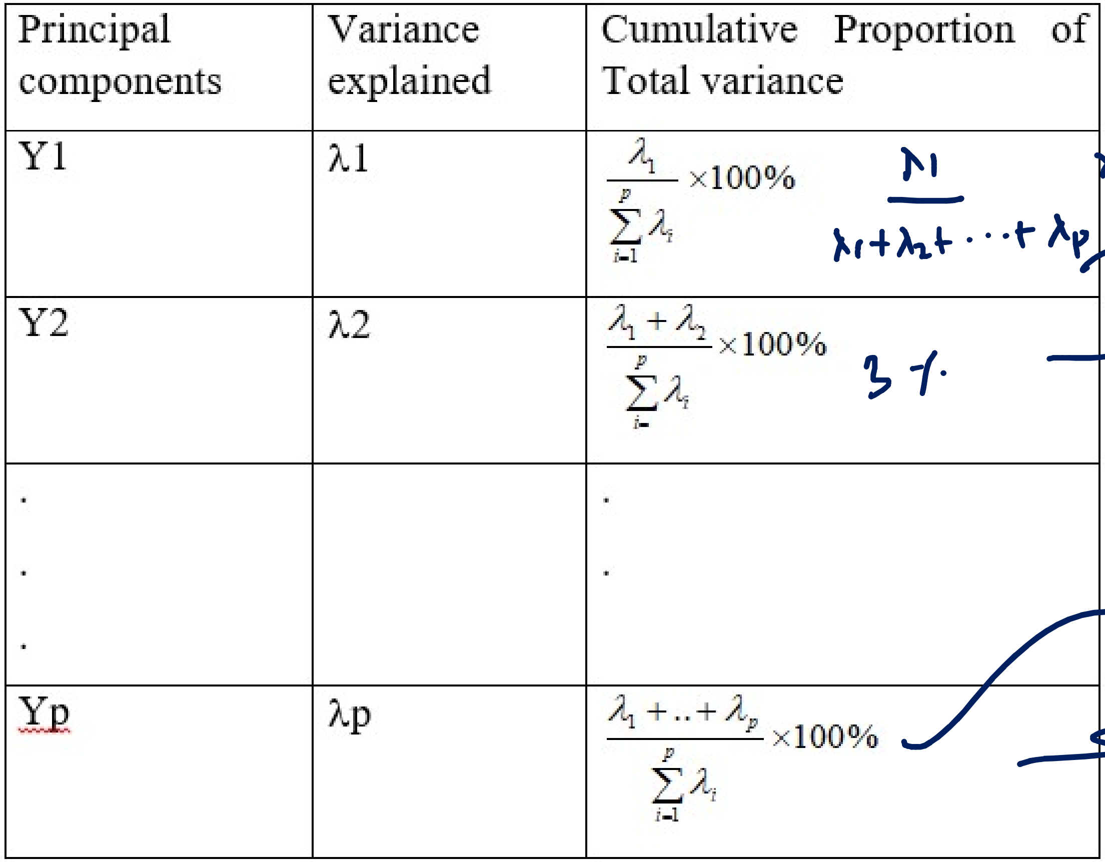
$Y_3 \neq Y_4$

$\vdots$

$Y_{p-1} \neq Y_p$

# How many principal components to retain?

a. **Retain sufficient components to account for a specified percentage of the total variance, say 80%.**



| Principal components | Variance explained | Cumulative Proportion of Total variance |
|---|---|---|
| Y1 | λ1 | $\frac{\lambda_1}{\sum_{i=1}^{p}\lambda_i}\times 100\%$ $\frac{\lambda_1}{\lambda_1+\lambda_2+\cdots+\lambda_p}$ |
| Y2 | λ2 | $\frac{\lambda_1+\lambda_2}{\sum_{i=}^{p}\lambda_i}\times 100\%$ 3% |
| .<br>.<br>. | | .<br>. |
| Yp | λp | $\frac{\lambda_1+..+\lambda_p}{\sum_{i=1}^{p}\lambda_i}\times 100\%$ |

$\times 100 = 89\%$

$89 + 3 = 92\%$

$\frac{\lambda p}{\Sigma \lambda_i} \times 100$

100%

## How many principal components to retain?

b. Exclude those components whose eigenvalues are less than the average of the eigenvalues

$$1, 2, 3 \longrightarrow \left[ \frac{\sum_{i=1}^{p} \lambda_i}{p} \right] \nrightarrow 4, 5, \ldots, p$$

# Another case of PCA

If the variables are measured on scales with widely different ranges or if the measurement units are not commensurate, then obtain principal components from correlation matrix. In this case principal components corresponding to S will be dominated by the variables with large variances. The other variables will contribute little. For a balanced representation in such cases, components of R (correlation matrix) may be used.

## Example:

**X1: Sales in the Rs. 10000 to Rs. 350000 range.**
**X2: Ratio in the 0.01 to 0.60 range.**

**In this case, we have to standardize the variable first. Let Z be the standardized variable, then**

$$Z=(z_1, z_2, \ldots\ldots\ldots, z_p)', \quad where \quad z_i = \frac{X_i - \bar{x}_i}{s_i}$$

$$R=\begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & \cdots & 1 \end{bmatrix}$$

**Where,**

$$r_{ij} \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{n} (x_{jk} - \bar{x}_j)^2}}$$

## Example:

**Let** $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq, \ldots\ldots, \hat{\lambda}_p$ **be the eigen values** $\underset{\sim}{e_1}, \underset{\sim}{e_2}, \ldots\ldots, \underset{\sim}{e_p}$ **be the corresponding eigenvectors.**

### Principal Components:

$$Y_1 = \underset{\sim}{e_1'} \underset{\sim}{Z} = e_{11}Z_1 + e_{21}Z_2 + \ldots\ldots + e_{p1}Z_p$$

$$Y_2 = \underset{\sim}{e_2'} \underset{\sim}{Z} = e_{12}Z_1 + e_{22}Z_2 + \ldots\ldots + e_{p2}Z_p$$

.

.

.

$$Y_p = \underset{\sim}{e_p'} \underset{\sim}{Z} = e_{1p}Z_1 + e_{2p}Z_2 + \ldots\ldots + e_{pp}Z_p$$

| PC | ve | cumulative prop^2 |
|---|---|---|
| 1 | $\lambda_1$ | $\frac{\lambda_1}{p} \times 100\%$ |
| 2 | $\lambda_2$ | $\frac{\lambda_1 + \lambda_2}{p} \times 100\%$ |

## Example:

## Example:

**Consider an example with 5 variables. The correlation matrix based on 100 observations is**

$$R = \begin{pmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1.000 & 0.599 & 0.389 & 0.322 \\ 0.509 & 0.599 & 1.000 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1.000 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1.000 \end{pmatrix}$$

**Eigenvalues and eigenvectors are**

$\hat{\lambda}_1 = 2.857$ $\quad$ $\hat{e}_1 = (0.464, 0.457, 0.470, 0.421, 0.421)'$ ✓

$\hat{\lambda}_2 = 0.809$ $\quad$ $\hat{e}_2 = (0.240, 0.509, 0.260, -0.526, -0.582)'$ ✓

$\hat{\lambda}_3 = 0.540$ $\quad$ $\hat{e}_3 = (-0.412, 0.178, 0.335, 0.541, -0.435)'$

## Example:

$\hat{\lambda}_4 = 0.452$ $\hat{e}_4 = (0.387, 0.206, -0.662, 0.472, -0.382)$

$\hat{\lambda}_5 = 0.343$ $\hat{e}_5 = (-0.451, 0.676, -0.4, -0.126, 0.385)$

Total sample variance explained by first two PC

$$= \frac{\lambda_1 + \lambda_2}{p} \times 100\% = \left(\frac{2.857 + 0.809}{5}\right) \times 100$$

$$= 73\%$$

## Example:

First two principal components :-

$$Y_1 = 0.464Z_1 + 0.457Z_2 + 0.470Z_3 + 0.421Z_4 + 0.421Z_5$$

$$Y_2 = 0.240Z_1 + 0.509Z_2 + 0.260Z_3 - 0.526Z_4 - 0.582Z_5$$

**Example:**

Apply PCA :-
$$\left(\begin{array}{cccccc} 19 & 22 & 6 & 2 & 2 & 20 \\ 12 & 6 & 9 & 15 & 13 & 5 \end{array}\right) \begin{array}{l} \to x_1 \\ \to x_2 \end{array}$$

Avg. of row 1 $= \bar{x}_1 = 12$

Avg. of row 2 $= \bar{x}_2 = 10$

Centered Matrix $(A) = \begin{pmatrix} 7 & 10 & -6 & -9 & -10 & 8 \\ 2 & -4 & -1 & 5 & 3 & -5 \end{pmatrix}$

Covariance Matrix $S = \dfrac{AA'}{n-1} = \dfrac{AA'}{5}$

## Example:

$$S = \frac{AA'}{5} = \underset{2\times 6}{\begin{pmatrix} 7 & 10 & -6 & -9 & -10 & 8 \\ 2 & -4 & -1 & 5 & 3 & -5 \end{pmatrix}} \underset{6\times 2}{\begin{bmatrix} 7 & 2 \\ 10 & -4 \\ -6 & -1 \\ -9 & 5 \\ -10 & 3 \\ 8 & -5 \end{bmatrix}}$$

$$= \frac{1}{5}\begin{pmatrix} 430 & -135 \\ -135 & 80 \end{pmatrix}$$

$$S = \begin{pmatrix} 86 & -27 \\ -27 & 16 \end{pmatrix}$$

$$|S - \lambda I| = \left| \begin{pmatrix} 86 & -27 \\ -27 & 16 \end{pmatrix} - \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right|$$

## Example:

To find eigen values of S,

$$\begin{vmatrix} 86-\lambda & -27 \\ -27 & 16-\lambda \end{vmatrix} = (86-\lambda)(16-\lambda)-729$$

$$= 1376 - 86\lambda - 16\lambda + \lambda^2 - 729$$

$$= \lambda^2 - 102\lambda + 647$$

$$\lambda = [95{\cdot}2,\ 6{\cdot}8] = [\lambda_1,\ \lambda_2]$$

## Example:

To find eigen vectors corresponding to eigen value,

for $\lambda_1 = 95.2$

$$(86-95.2)x - 27y = 0$$

$$-9.2x - 27y = 0$$

$$x = -\frac{27}{9.2}y$$

$y=1$, $x = -2.9347$

$$\sqrt{x^2+y^2} = \sqrt{(-0.2934)^2+1^2} = 3.1$$

normalising $(x,y) \Rightarrow \left(\frac{-2.9347}{3.1}, \frac{1}{3.1}\right)$

we get the direction as

$$(-0.95, 0.32) = \hat{e}_1$$

## Example:

$|||^{ly}$, we can find eigen vector corresponding to $\lambda_2 = 6.8$,

$$79.2x - 27y = 0$$

$$79.2x = 27y$$

$$x = \frac{27}{79.2} y$$

let $y = 1$, $x = 0.34091$

Normalising (dividing by 1.0565)

$$\hat{e}_2 = (0.32, 0.95)$$

## Example:

The first PC accounts for 95.2 % of total sample variance

So, we can compute PC as

✓ $Z_1 = -0.95X + 0.22Y$

✓ $Z_2 = 0.32X + 0.95Y$

$$P = AU = \begin{bmatrix} 7 & 2 \\ 10 & -4 \\ -6 & -1 \\ -9 & 5 \\ -10 & 3 \\ 8 & -5 \end{bmatrix}_{6\times 2} \begin{bmatrix} -0.95 & 0.32 \\ 0.32 & 0.95 \end{bmatrix}_{2\times 2} = \begin{bmatrix} -6.01 & 4.14 \\ -10.78 & -0.6 \\ 5.38 & -2.87 \\ 10.15 & 1.87 \\ 10.46 & -035 \\ -9.2 & -2.19 \end{bmatrix}$$

## Example:

2) $$Z' = \begin{bmatrix} 2.5 & 0.5 & 2.2 & 1.9 & 3.1 & 2.3 & 2 & 1 & 1.5 & 1.2 \\ 2.4 & 0.7 & 2.9 & 2.2 & 3 & 2.7 & 1.6 & 1.1 & 1.6 & 0.9 \end{bmatrix}$$

$\rightarrow$ $\bar{x}_1 = 1.81$ , $\bar{x}_2 = 1.91$

$$A' = \begin{bmatrix} 0.69 & -1.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.91 & -0.31 & -0.71 \\ 0.49 & -1.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -1.01 \end{bmatrix}$$

$$S = \frac{A'A}{n-1} = \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$$

## Example:

To find eigen values & vectors

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

$$\Rightarrow \lambda_1 = 1.2840 \qquad \lambda_2 = 0.0490$$

Eigen vectors

for $\lambda_1$, $\hat{e}_1 = [0.67787, 0.73517]'$

for $\lambda_2$: $\hat{e}_2 = [-0.7351, 0.67787]'$

## Example:

$$U = \begin{bmatrix} 0.67787 & -0.7351 \\ 0.7351 & 0.6778 \end{bmatrix}$$

variance explained by 1st $PC = \dfrac{\lambda_1}{\lambda_1 + \lambda_2} \times 100\%$

$$= \frac{1.2840}{1.2840 + 0.0490} = 96.32\%$$

⟶ 2nd $PC = 3.68\%$

## Example:

$$z_1 = 0.67787x_1 + 0.73517x_2$$

$$z_2 = -0.73517x_1 + 0.67787x_2$$

$$P = AU =$$

## Example:

## Example:

## Derive the new data:

$\overset{P}{\text{FinalData}}$ = RowFeatureVector x RowZeroMeanData

$\overset{U}{\textit{RowFeatureVector}}$ is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top.

$\overset{A}{\textit{RowZeroMeanData}}$ is the mean-adjusted data transposed, i.e., the data items are in each column, with each row holding a separate dimension.

## Reconstruction of Original Data:

**We can reconstruct the original data with respect to the principal components chosen as:**

1. **FinalData = RowFeatureVector x RowZeroMeanData** $\Rightarrow P = AU$
2. **RowZeroMeanData = RowFeatureVector-1 x FinalData**
3. **RowOriginalData = (RowFeatureVector-1 x FinalData) + OriginalMean**

**Note:**

- **As we use unit eigen vectors, the inverse is same as the transpose.**
- **If we reduce the dimensionality (i.e., r < m), obviously, when reconstructing the data we lose those dimensions we chose to discard.**
- **In our example if we considered only a single eigenvector, then final data is Z1 only and the reconstruction yields the data where the variation along the principal component 1 is preserved while the variation along the other component will be lost.**

## Example:

**Example: Apply the principal component analysis for the data given below:**

$$\mathbf{Y} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix}$$

## Example: