Submitted by-Aditya Gautam

Roll No-12

M.sc. 3rd semester

Date of Assignment-17/12/2020

Date of Submission-05/01/2021

Experiment No-11

Topic- Principal component analysis.

Problem- The following values are obtained from a sample drawn from a 5-variate distribution:

$\underset{\sim}{\bar{X}}' = (0.0054 \quad 0.0048 \quad 0.0057 \quad 0.0063 \quad 0.0037)$

$$R = \begin{pmatrix} 1 & 0.577 & 0.500 & 0.387 & 0.462 \\ & 1 & 0.599 & 0.389 & 0.322 \\ & & 1 & 0.436 & 0.426 \\ & & & 1 & 0.523 \\ & & & & 1 \end{pmatrix}$$

Carry out the PCA and draw your conclusion from scree plot. Also, find the values of $\rho_{y_1 z_1}$ and $\rho_{y_2 z_2}$, where Y is the i^{th} sample principal component and Z_i is the standardised variable corresponding to X_i.

Theory-

Suppose, we have the data on p-variables $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_p$ and $\underset{\sim}{x}' = (x_1, x_2, \ldots . x_p)$ is random vector from the same p-dimensional population with mean vector $\underset{\sim}{\mu}$ and variance covariance matrix $\sum$.

The correlation matrix R is the covariance matrix of the standardized variables

$$Z_1 = \frac{x_1 - \mu_1}{\sqrt{\sigma_p^2}}, Z_2 = \frac{x_2 - \mu_2}{\sqrt{\sigma_p^2}}, \ldots, Z_p = \frac{x_p - \mu_p}{\sqrt{\sigma_p^2}}$$

The i^{th} principal component of the standardized variables is given by -

$$\underset{\sim}{y}_i = \underset{\sim}{\hat{e}}_i \underset{\sim}{Z} = (\hat{e}_{i1}, \hat{e}_{i2}, \ldots . . \hat{e}_{ip}) \begin{pmatrix} Z_1 \\ Z_2 \\ . \\ . \\ . \\ Z_p \end{pmatrix}$$

$= \hat{e}_{i1} Z_1 + \hat{e}_{i2} Z_2 + \cdots . . \hat{e}_{ip} Z_p$; i=1, 2, …, p

Where, $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \ldots . . , (\hat{\lambda}_p, \hat{e}_p)$ are the eigen value-eigen vector pairs of the correlation matrix R.

A useful aid to determine principal components to retain is a scree plot. With the Eigen values ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus i, i.e the magnitude of the Eigen value versus its number.

To determine the appropriate number of components we look for an elbow (or bend) in the scree plot. The number of components is taken to be the point at which the remaining Eigen values are relatively small and all about the same value.

The correlation between $\underset{\sim}{Y}_1$ and Z_1 and $\underset{\sim}{Y}_2$ and Z_2 is given by -

$$\rho_{y_1 z_1} = e_{11} \sqrt{\lambda_1}$$

$$\rho_{y_2 z_2} = e_{22} \sqrt{\lambda_2}$$

Calculation-

The R-programming to obtain the solution for the given problem-

```
library('ggplot2')

R=array(c(1,0.577,0.500,0.387,0.462,0.577,1,0.599,0.389,0.322,0.5,0.599,1,0.436,0.426,0.387,0.
389,0.436,1,0.523,0.462,0.322,0.426,0.523,1),dim=c(5,5))

R

e_val=eigen(R)$value

e_val

e_vec= eigen(R)$vectors

e_vec

p=mat.or.vec(5,1)

for (i in 1:5){

p[i]=e_val[i]/5}

p[i]

prop=cumsum(p)

prop

ry1_z1=e_vec[1,1]*sqrt(e_val[1])

ry1_z1

ry1_z2=e_vec[2,2]*sqrt(e_val[2])

ry1_z2

i=c(1,2,3,4,5)

i

Table = data.frame(i,e_val)

Table

View(Table)
```

```
plot(i,e_val,type="o",main="Scree Plot")

#using ggplot we get required graph

ggp = ggplot(data=Table,mapping=aes(x=i,y=e_val))+geom_point()+geom_line()

labs(

  title = paste("Scree Plot")

 )

ggp
```

Conclusion-

Proportion of variation explained by first sample principal component

$$= \frac{\lambda_1}{\sum_{i=1}^{n} \hat{\lambda}_i} = \frac{2.8527958}{5} = 0.5706$$

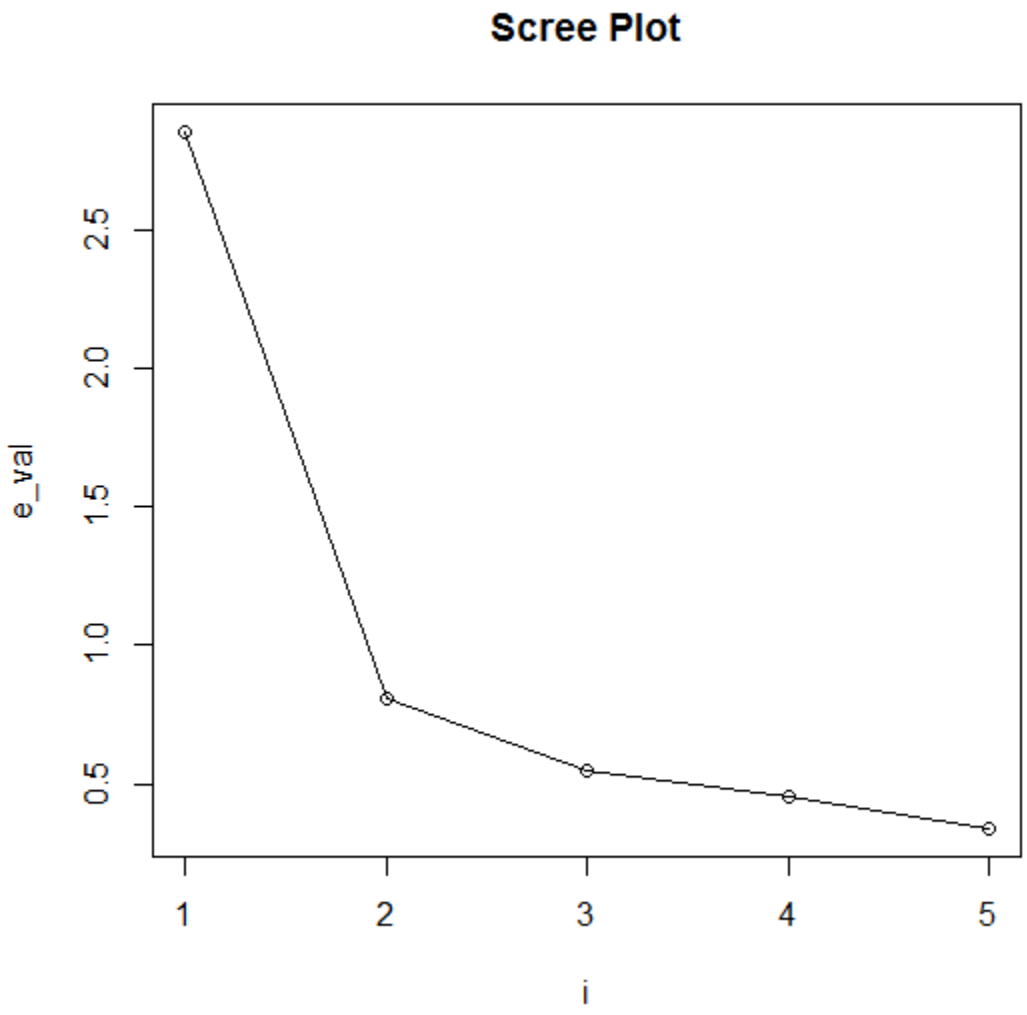
Proportion of variation explained by second sample principal component

$$= \frac{\lambda_2}{\sum_{i=1}^{n} \hat{\lambda}_i} = \frac{0.8080454}{5} = 0.1616$$

Proportion of variation explained by third sample principal component

$$= \frac{\lambda_3}{\sum_{i=1}^{n} \hat{\lambda}_i} = \frac{0.5436298}{5} = 0.1087$$

Also, from the Scree plot, we see that a bend occurs corresponding to the Eigen value 3 which is shown in the graph below-



Since the first three sample pc's can very well summarize the total sample variance about 84% of the total variance, a reduction in the data of 5 variables to 3 principal components is reasonable. Therefore, the number of principal components to be retained is 3 and the principal components are ---

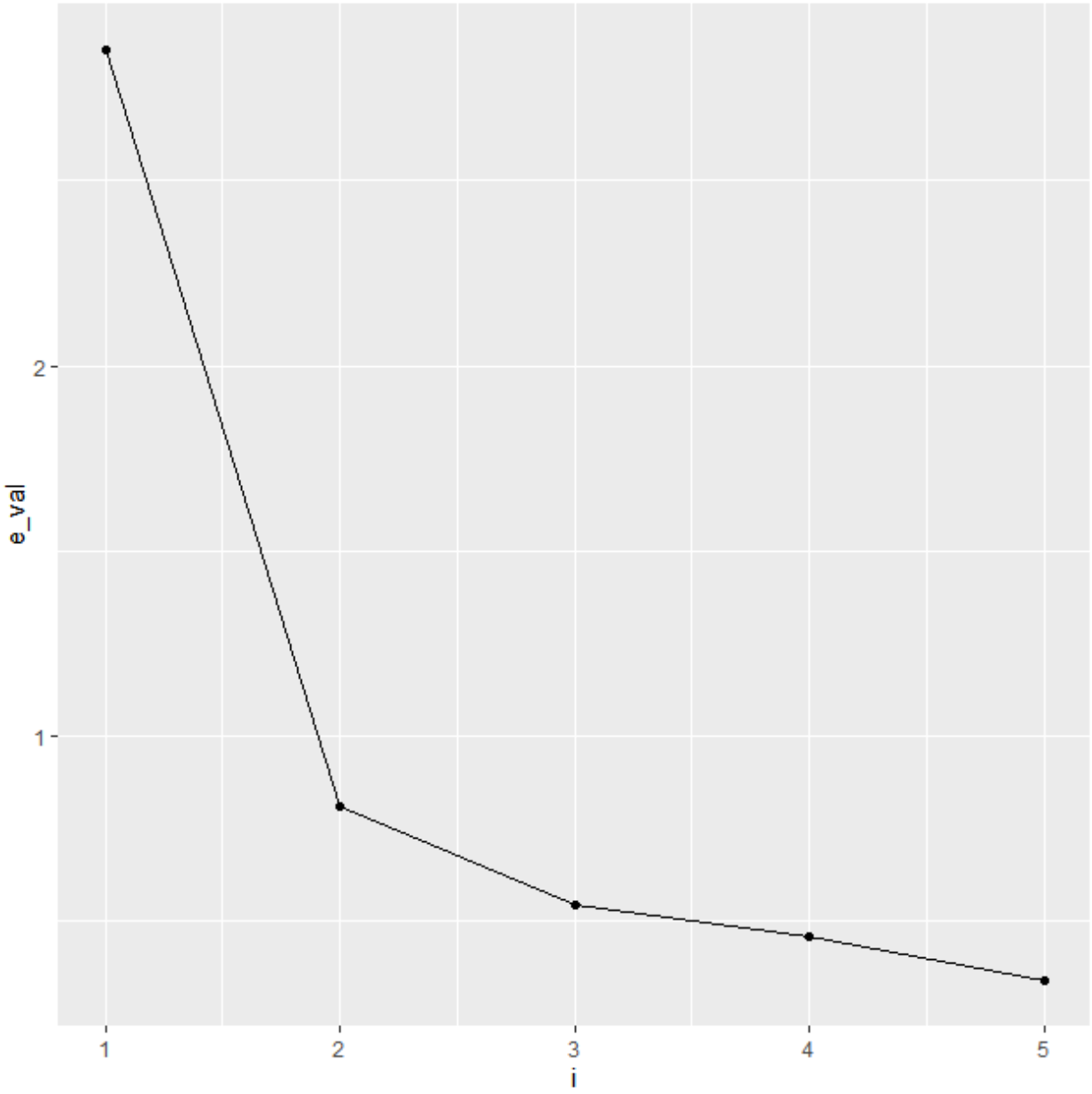
$$y_1 = (-0.4625824 \ \ -0.4577342 \ -0.4691988 \ -0.4222471 \ \ -0.4219666) \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \end{pmatrix}$$

$$y_2 = (0.2378984 \;\; 0.5133099 \;\; 0.259208 \;\; -0.5238355 \;\; -.5816568)\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \end{pmatrix}$$

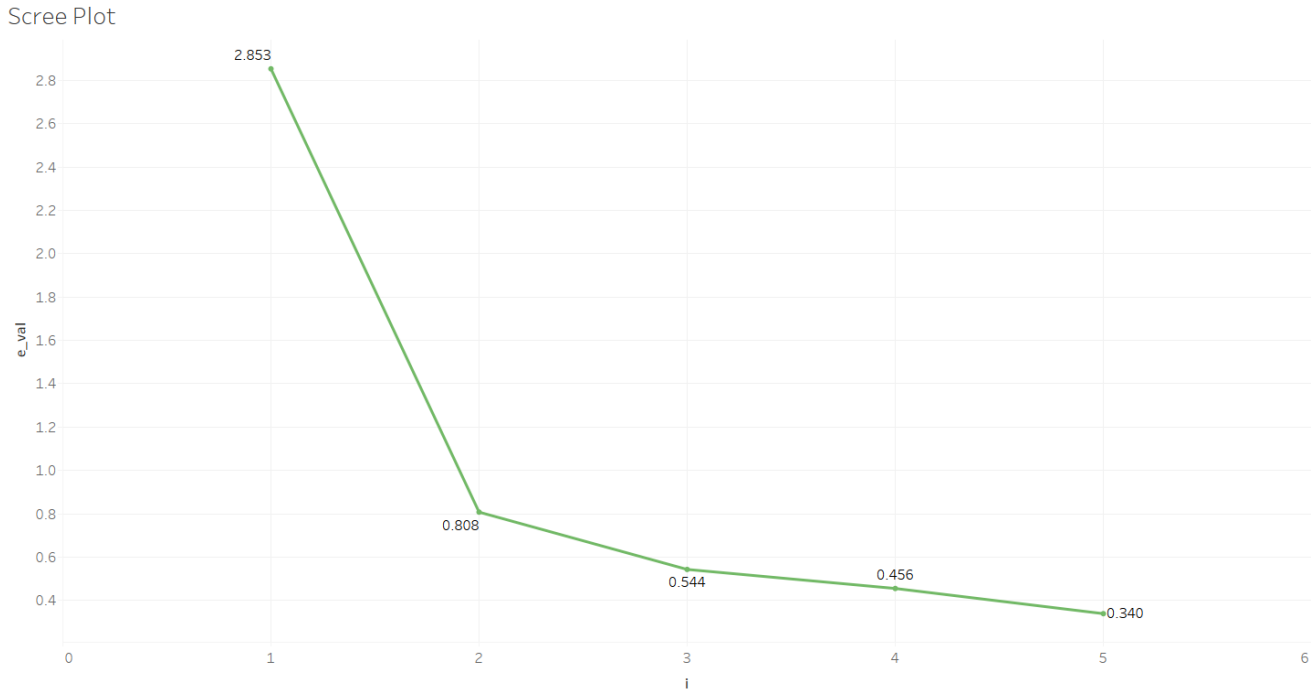
$$y_3 = (0.6311136 \;\; -0.1625001 \;\; -0.3715296 \;\; -0.5158427 \;\; 0.4137166)\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \end{pmatrix}$$

Finally, $\rho_{y_1 z_1}$ = -0.7813119 and $\rho_{y_2 z_2}$ = 0.4614211.

Using GGPLOT we can plot the required graph-



Using Tableau we plot the required graph



I vs. e_val. The marks are labeled by sum of e_val.