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:

= (0.0054 0.0048 0.0057 0.0063 0.0037)



Carry out the PCA and draw your conclusion from scree plot. Also, find the values of and, where Y is the sample principal component and is the standardised variable corresponding to .

**Theory-**

Suppose, we have the data on p-variables and x̰′= (random vector from the same p-dimensional population with mean vector and variance covariance matrix ∑.

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

The principal component of the standardized variables is given by -

=

= ; i=1, 2, …, p

Where, ) 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 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 and and and is given by -





**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-**

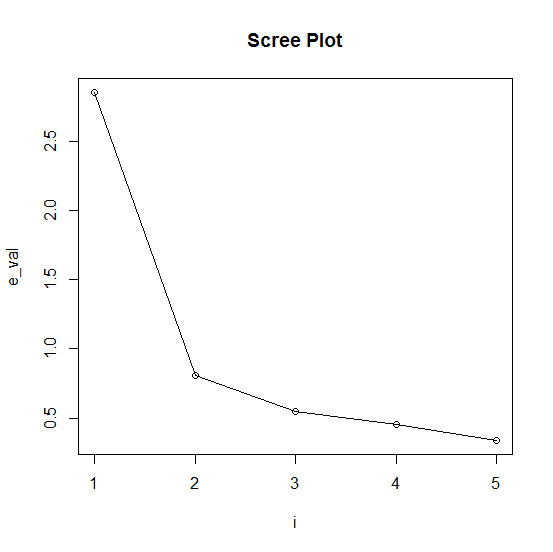
Proportion of variation explained by first sample principal component



Proportion of variation explained by second sample principal component

Proportion of variation explained by third sample principal component

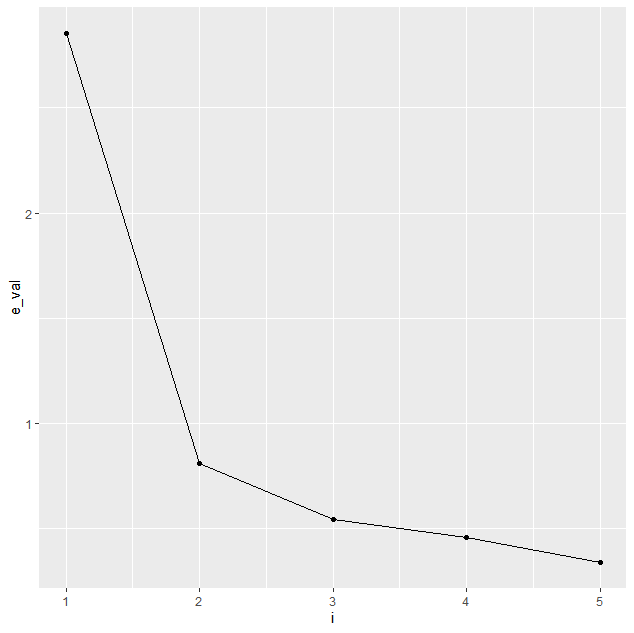
Also, from the Scree plot, we see that a bend occurs corresponding to the Eigen value 3 which is shown in the graph below-



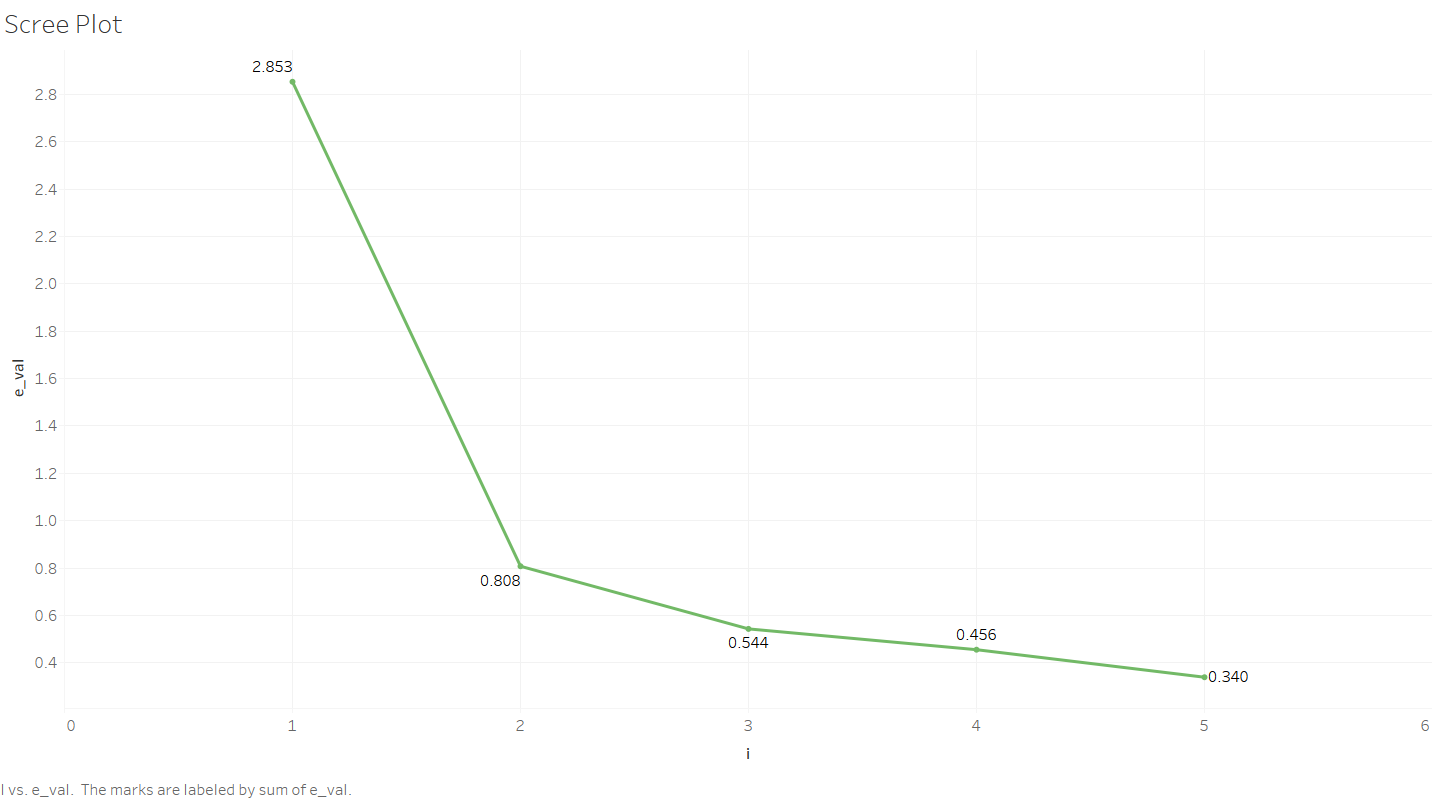
Since the first three sample pc’s can very well summarize the total sample variance about 84% of the total variance, a reduction in the data of 5 variables to 3 principal components is reasonable. Therefore, the number of principal components to be retained is 3 and the principal components are ---

Finally, = -0.7813119 and= 0.4614211.

**Using GGPLOT we can plot the required graph-**

****

**Using Tableau we plot the required graph**

****