Maitreya Kanithar
BE-IT 8084

MLA Assignment 2

Title :-
Principal Component Analysis

Aim:-
Finding principal components, variance and standard deviation calculations of principal components.

Theory:-

q1) Explain dimensionality reduction.

→ Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal values.

There are two components of dimensionality reduction-

a) Feature selection-
In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem.

b) Feature extraction-
This reduces the data in a high dimensional space to a lower dimension space.

Dimensionality reduction may be both linear or non-linear depending upon the method used.

The various methods used for dimensionality reduction include-

a) Principal Component Analysis (PCA)
b) Linear Discriminant Analysis (LDA)
c) Generalized Discriminant Analysis (GDA)

Q2) Explain Forward selection and Backward selection methods.

→ Forward selection-
Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step you add the one variable that gives the single best improvement to your model.

It typically begins with only an intercept. One tests the various variables that may be relevant, and the 'best' variable is added to the model. As the model continues to improve we continue the process, adding one variable at a time and testing at each step. Once the model no longer improves with adding more variables, the process stops.

Backward selection-
In backward selection you start with a model that includes every possible variable and eliminate the extraneous variables one by one.

In each step feature is removed one by one and the error is checked. Process continues until significant decrease in error occurs.

Q3) Differentiate between Feature selection and Feature extraction.

->

	Feature selection	Feature extraction
a)	It finds a smaller subset of n-dimensional data set to create a data model.	It involves transforming n dimensional data into spaces of fewer dimensions.
b)	It is the technique of finding k features of the d dimension that gives us the most information and discard the other (d-k) dimensions.	It is a technique of finding a new set of k dimensions that are combinations of the original d dimensions.
c)	Subset selection is one of the widely used method for feature selection method.	The best known and most widely used feature extraction method is Principal Component Analysis.

Q4) Explain Principal Component Analysis.

-> Principal Component Analysis.

Principal Component Analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large datasets, by transforming a large set of variables into smaller ones that still contain most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller datasets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

Steps of PCA-

i) Standardization
ii) Covariance matrix computation
iii) Compute eigen values and eigen vectors of the covariance matrix to identify the principal components

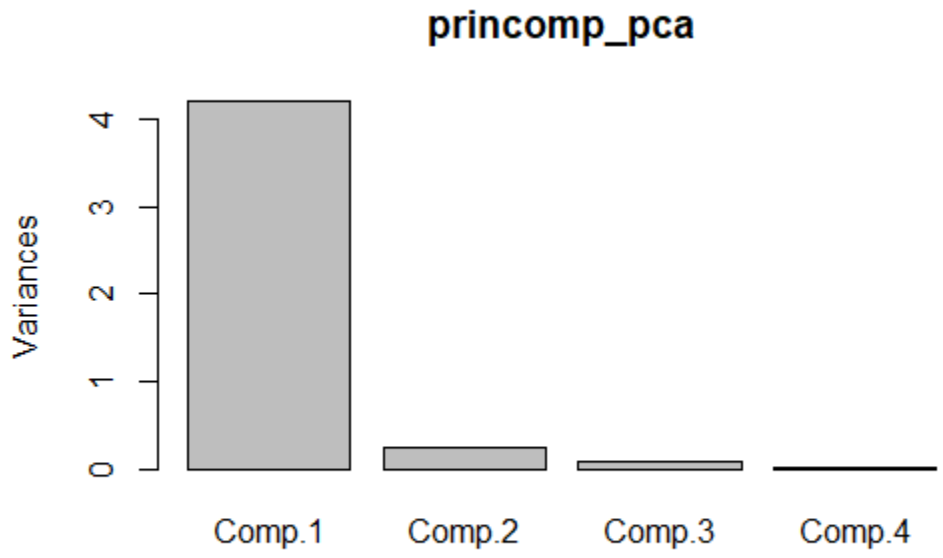
INPUT:-

```
#loading iris dataset
data(iris)
summary(iris)


#covariance matrix
cov_matrix=cov(iris[1:4])
cov_matrix


#eigen values and eigen vectors
cov_eigen=eigen(cov_matrix)
cov_eigen


#princomp and prcomp
princomp_pca=princomp(iris[1:4])
summary(princomp_pca)

prcomp_pca=prcomp(iris[1:4])
summary(prcomp_pca)

# Comparing variance values
cov_eigen$values
princomp_pca$sdev^2
prcomp_pca$sdev^2

#plotting the data(princomp)
plot(princomp_pca)
screeplot(princomp_pca, type = "lines")
biplot(princomp_pca)

#plotting the data(prcomp)
plot(prcomp_pca)
screeplot(prcomp_pca, type = "lines")
biplot(prcomp_pca)
```

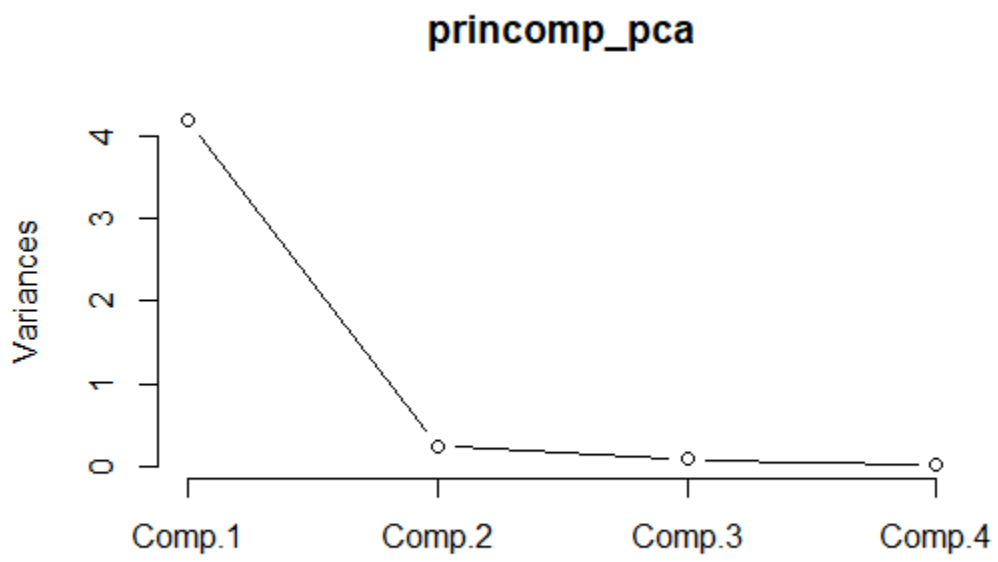
OUTPUT:-

```
> #loading iris dataset
> data(iris)
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
>
> #covariance matrix
> cov_matrix=cov(iris[1:4])
> cov_matrix
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
>
>
> #eigen values and eigen vectors
> cov_eigen=eigen(cov_matrix)
> cov_eigen
eigen() decomposition
$values
[1] 4.22824171 0.24267075 0.07820950 0.02383509

$vectors
           [,1]       [,2]       [,3]       [,4]
[1,]  0.36138659 -0.65658877 -0.58202985  0.3154872
[2,] -0.08452251 -0.73016143  0.59791083 -0.3197231
[3,]  0.85667061  0.17337266  0.07623608 -0.4798390
[4,]  0.35828920  0.07548102  0.54583143  0.7536574

>
>
> #princomp and prcomp
> princomp_pca=princomp(iris[1:4])
> summary(princomp_pca)
Importance of components:
```

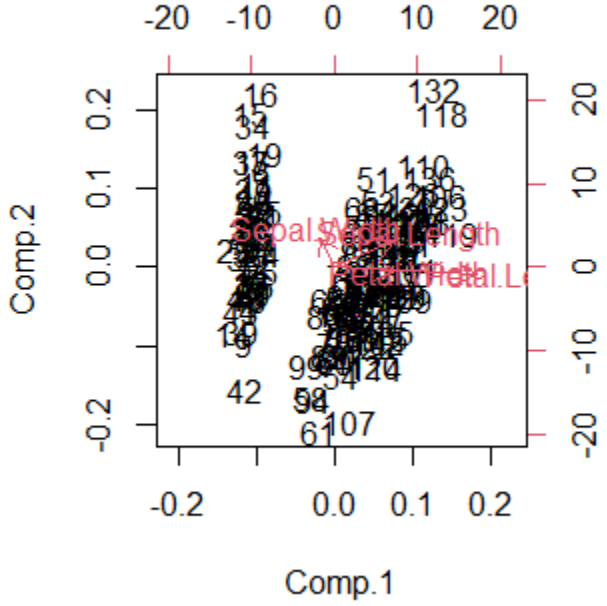
```
                       Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
>
> prcomp_pca=prcomp(iris[1:4])
> summary(prcomp_pca)
Importance of components:
                          PC1     PC2    PC3     PC4
Standard deviation     2.0563 0.49262 0.2797 0.15439
Proportion of Variance 0.9246 0.05307 0.0171 0.00521
Cumulative Proportion  0.9246 0.97769 0.9948 1.00000
>
> # Comparing variance values
> cov_eigen$values
[1] 4.22824171 0.24267075 0.07820950 0.02383509
> princomp_pca$sdev^2
    Comp.1     Comp.2     Comp.3     Comp.4
4.20005343 0.24105294 0.07768810 0.02367619
> prcomp_pca$sdev^2
[1] 4.22824171 0.24267075 0.07820950 0.02383509
>
> #plotting the data(princomp)
> plot(princomp_pca)
```



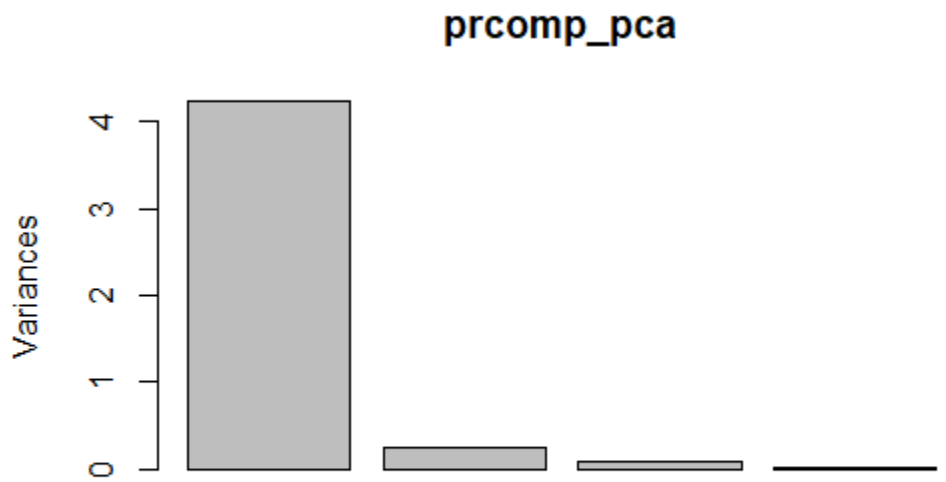
> screeplot(princomp_pca, type = "lines")



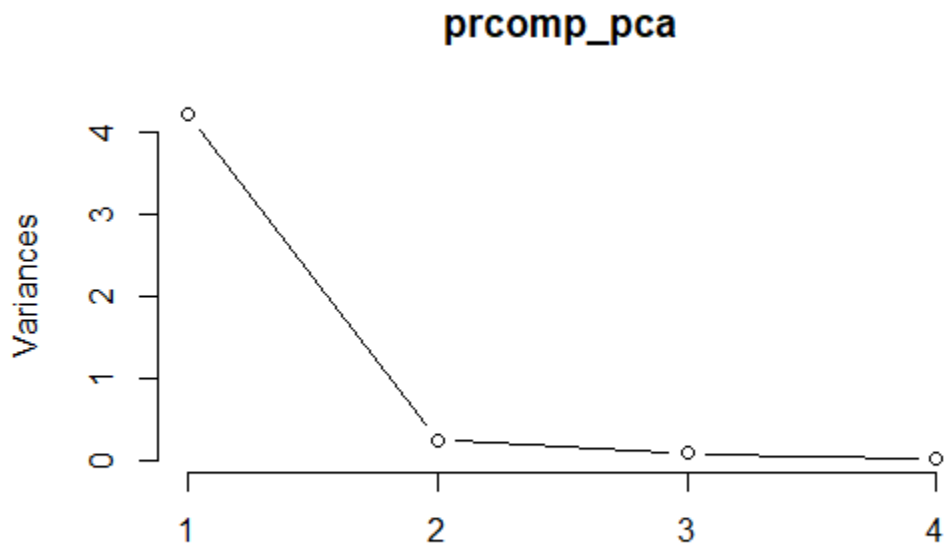
> biplot(princomp_pca)



```
> #plotting the data(prcomp)
> plot(prcomp_pca)
```




```
> screeplot(prcomp_pca, type = "lines")
```



```
> biplot(prcomp_pca)
```