IRIS DATA SET ANALYSIS

USING PRINCIPAL COMPONENT ANALYSIS

Shyam Modi

***Abstract :*** When there are several parameters in a particular data set, the accuracy of its model reduces. It is known as the “Curse of dimensionality”. Principal Component Analysis is applied to increase the accuracy of a particular data model. In this paper, I took a set of data called “Iris” which already exists in RStudio. Iris data consists of 150 observations and 4 features. Libraries that I used are biplot and screenplot. Using the function princomp(), I proceeded to do the PCA. After its execution, I found the PCA object using the function summary(). Finally, I plotted the components of PCA to give a much better visual representation to look at the data.

***Keywords : Principal Component Analysis, Iris data set, R programming, data analysis.***

1. **Introduction *:*** Principal Component Analysis is a technique which uses statistics to emphasizes variation and finds strong patterns from a set of data of interrelated variables. It makes the data easy to inspect and visualize. It works as a dimension reducer by reducing the number of parameters to improve model accuracy. It reduces a big amount of data set to a smaller data set while still retaining most of the information. Variable reduction makes it simpler for us interpret and analyze the data. The princomp() function measures the principal component of any data. Then I will compare its values with eigenvalues as well.

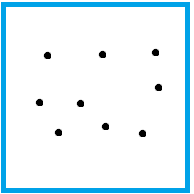
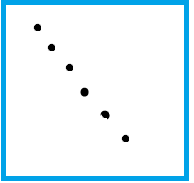
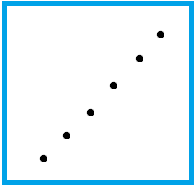
***Few Important Terms***

Variance- It measures the variability or change in the data set. It is calculated by average squared deviation upon mean score.

var(x) = [∑(xi - ͞x)^2] ÷ N

Covariance- It is the calculation of joint variability of two random variables. When corresponding data values from two sets move in similar direction.

cov(x,y) = [∑(xi - ͞x)(yi - ͞y)] ÷ N

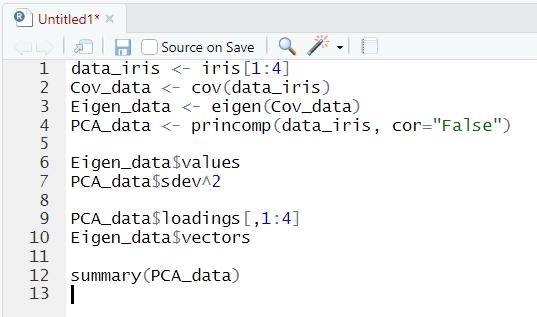
Near zero co-variance Negative co-variance Positive co-variance

When x and y are values are in increasing order, they are in a positive covariance.

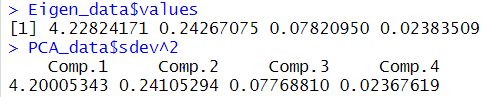
When they are in decreasing order, they are in negative covariance.

1. **Proposed System Approach**

Now I will show how to implement PCA using RStudio. The princomp() function measures the components of the data.Then I calculated eigenvalues separately.

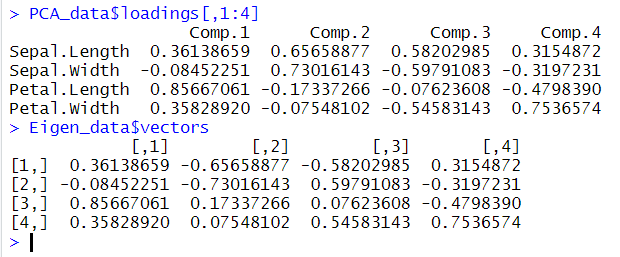


1. In the first step we called the iris set, and took 4 features from the set because the 5th feature is categorical data.
2. cov() measures the covariance matrix of Iris.
3. Next I calculated the eigenvalues.
4. Then I used princomp() function to find the principal components.
5. Then I compared the result of PCA and eigenvalues.



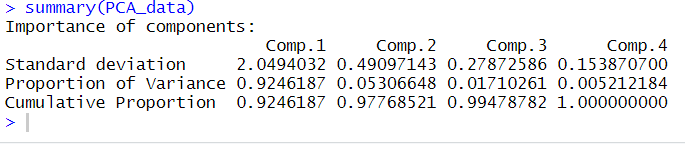
The values are pretty similar.

1. Then I compared the output of eigenvectors and loadings of both models.



The values on both eigenvectors and loadings are the same. Here, the sign of the values are random and meaningless. The sign denotes the direction of that given variable in that Principal Component.

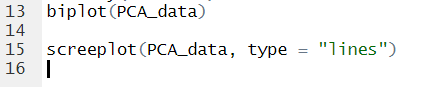
1. After that, I viewed the summary.

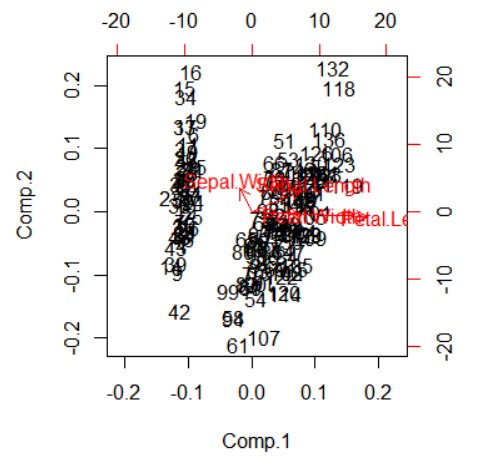


To improve accuracy, I took first two components together and obtained cumulative frequency of 97.7%.

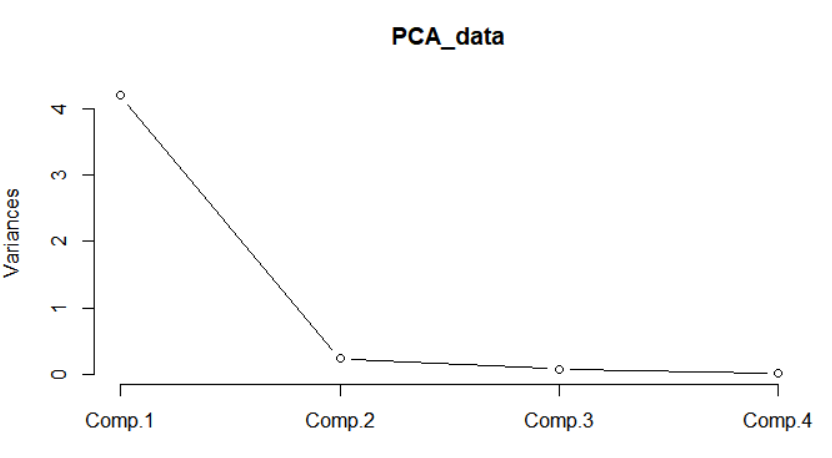
I plotted the data as well:-

1. Biplot.





The figure above is a biplot of the two components



The figure above is screenplot of the two components.

1. **Future Scope and Discussion**

This method can be applied in many situations. More versions or software of PCA can be made to make faster, robust and more efficient. New features can be added to make it more advanced. Due to this, it is also used in image compression to genetic analysis.

1. **Conclusion**

PCA is very wide topic that is growing in today’s world. It is helps in improve model accuracy, makes the data easy to visualize and interpret. It also assists in show regular patterns in the data. Unfortunately, one of the limitations of PCA is that it heavily depends upon variance and mean of the data. If the parameters of the data are initially orthogonal and not related to each other, PCA will be unable to give any useful output. Instead, it will simply order the parameters in declining values of variances.