Principal Component Analysis - Online Statistical Analysis Tool

O. P. Sheoran, Vinay Kumar, Hemant Poonia, Komal Malik

Abstract: ***An online module to deal with PCA has been developed in ASP scripting language based on Server-Client Architecture. The module produces descriptive statistics via sub-program Descriptive Stats, computes eigenvalues and eigenvector using MxEigen Jacobisub-program, order eigenvector through MxEigsrtsub-program and finally produces eigenvalues, eigenvectors, output loadings and components scores through Output Eigenval, Output Loadings, Output Scoressub-programs. A user friendly interface has been developed for entering or pasting the data, entering various parameters such as number of variables, number of observations and selection of covariance/correlation matrix. A complete procedure for how to perform principal component has also been provided in help file.***

Keywords: Principal Component, Eigenvalues. Eigenvectors, Component Scores, Loadings.

I. INTRODUCTION

Principal Component Analysis (PCA) is multivariate statistical tool used to analyze multidimensional data. PCA is used in almost all areas of research for manipulating large numbers of variables/attributes simultaneously and is helpful for retrieving important information from a complex data set. It reduces the dimensionality of a data set containing large number of interrelated variables and retains as much of the variation present in the original data set as possible. This is achieved by transforming the original set of variables into a smaller set of variables. These new variables correspond to linear combination of original variables and are called as principal components (PCs). The derived Principal components are uncorrelated and ordered so that the first few contain most of the variation present in the original variables. Main goals of PCA are to identify hidden pattern in the data set, to reduce the dimensionality of the data by removing noise and redundancy in the data and to identify correlated variables.

The PCA is mainly based on eigenvalues and eigenvectors of correlation/covariance matrices. The computation of eigenvalues and eigenvector require intensive matrix manipulation that are almost impossible by hand or desk calculator. Commercial statistical packages like SPSS, SAS, R and Excel Addins have been providing the facilities for PCA but user has to purchase software and proper training to install and use software is required.

Revised Manuscript Received on February 15, 2020.

O.P. Sheoran, Professor, Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar.
Email:opsheoran1968@gmail.com

Vinay Kumar, Assistant Scientist, [3]Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar.
Email:vinay.stat@gmail.com

Hemanat Poonia, Assistant Professor, [3]Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar.
Email:pooniahemant80@gmail.com

Komal Malik, Assistant Professor, Economics,,Govt.College, Nalwa-Hisar, Email: komalmalikeco@gmail.com

Keeping this in mind an attempt has been made to develop an online principal component analysis tool for the researchers who are unable to buy costly software. This tool is freely available at http://14.139.232.166/pca/pca.html.

II. MATERIAL AND METHODS

The online tool has been developed to provide significant features of principal component analysis including correlation matrix, eigenvalues, eigenvectors and principal components scores. The procedure consists of following basic steps:

1. **Center and Scale of Data:** The raw data is centered by subtracting mean from each variable. If the variances of the variables in data set are significantly different, then data is scaled to unit variance by dividing each variable by its standard deviation.
2. **Computation of covariance/correlation matrix**: Compute covariance matrix if the variables are measured on same units of measurements otherwise correlation matrix is computed and will be used for further analysis.
3. **Computation of the eigenvectors and the eigenvalues**: Computeeigenvalues and corresponding eighenvectors of correlation/covariance. Eigenvectors are ordered by eigenvalues from the highest to the lowest. The number of chosen eigenvectors will be the number of dimensions of the new data set.
4. **Compute transformed values:**Compute the principal components scores for selected PC's that can be used for further analysis.

The procedure is also depicted through flow-chart (Figure 1)

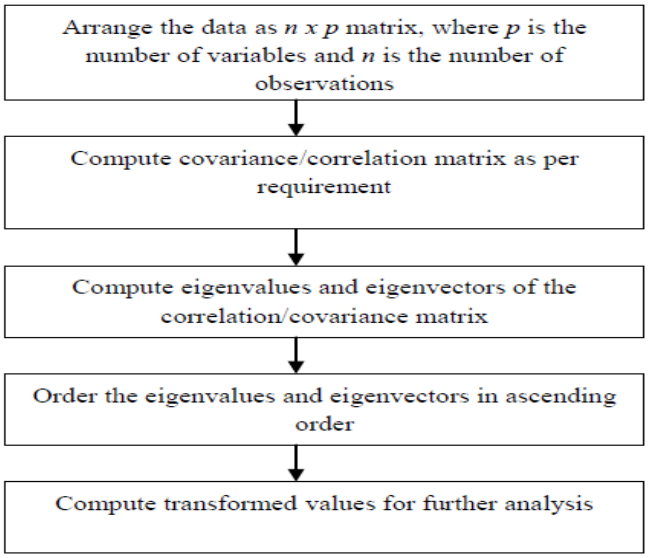


Figure 1: Flow chart of the basic steps of PCA.

III. ALGORITHM DESIGN

The algorithms for online module has been developed in ASP scripting language based on Server-Client Architecture. The tool consists of data entry page written in HTML to facilitate the user for entering raw data or paste the data in the text area provided. The data should be arranged in *n x p* matrix form where n represents number of observations and p represents number of variables. The data should be tab or space delimited. The variables names can be entered in another text area row-wise.On submitting the data for analysis new web page opens up asking for number of variables and number of observations per variable along with an option for PCA based on correlation or covariance matrix. The procedure consists of several sub-programs which are described as under:

DescriptiveStats(datamatrix, n, p, mean, V, sd, Cov, Corr) computes the descriptive statistics *viz* mean, variance, standard deviation, covariance and correlation matrices.Here input arguments are*datamatrix*, *n* and *p*where *Datamatrix* is raw data arranged in *n x p* matrix,*n*is the number of observations and*p* is number of variables whereas *V, Sd, Cov* and *Corr* are output arguments where*V*is Variances, *Sd*is Standard deviation, *Cov* is Covariance matrix and *Corr* is the Correlation matrix.

MxEigenJacobi (Cov, p, Eigen Vect, Eigen Val) and ***Mx EigenJacobi (Corr, p, Eigen Vect, EigenVal) c***omputesthe eigenvalues and eigenvectors form covarianceandcorrelation matrix obtained from*DestriptiveStat,* respectively depending upon the fact that you are using covariances or correlations for computation of eigenvalues and eigenvectors. The description of arguments of these sub-modules is given as under:

Cov/Corr – input argument (Covariance or Correlation matix)*p* – order of correlation/covariance matrix (input argument)

EigenVect– EigenVector (output argument)

EigenVal– EigenValues (output argument)

The Eigenvectors are ordered by eigenvalues from the highest to the lowest using the sub-program ***MxEigsrt (EigenVect, EigenVal, p)*** having following arguments

EigenVect– EigenVector (input/output argument)

EigenVal– EigenValues (input/output argument)

p – order of correlation/covariance matrix (input argument)

The results computed from sub-programs mentioned above are displayed in html format on new web page. The description of each sub program is given below

OutputEigenval(Mtrxtype, EigenVect, n, p, rf) - displays eigenvalues and eigenvectors

OutputLoadings(Mtrxtype, sd, EigenVect, EigenVal, p, rf, VariableName) – displays loadings

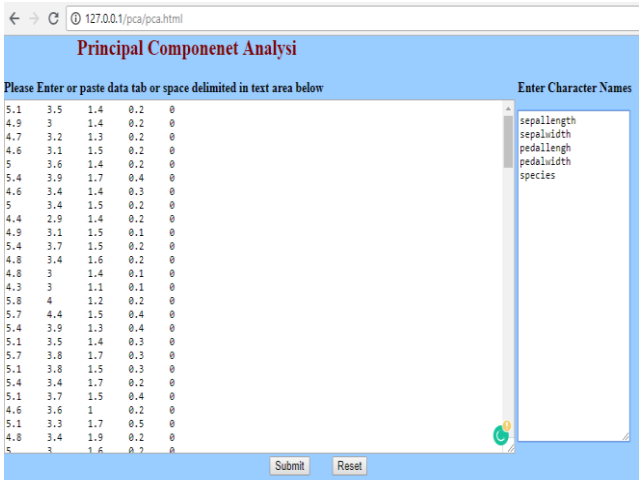
OutputScores(Mtrxtype, data, mean, sd, EigenVect, EigenVal, n, p, rf, CasesLabel) – displays the component scores

A. Validation of results:

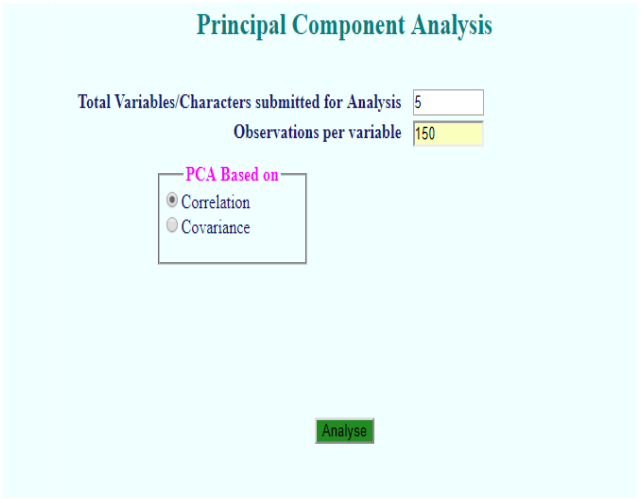
The output of module has been validated through iris data on sepal length, sepal width, petal length, petal width and spices on 150 iris flowers from *wikipedia. org/wiki/Iris_flower_data_set*..The results obtained from the modules are in agreement with standard packages like SPSS and R,

B. Procedure for Analysis of data

Enter or paste the data for principal component analysis in the text area provided in web page of the module under the heading **'Please Enter or paste data tab or space delimited in text area below'**. The data should be arranged in such a way that the first observations of all the variables/characters must be entered in first line and delimited by space or tab. Likewise, enter the data for all the observations of all the characters in subsequent lines. Make sure not to enter character/variables name in this text area. Character/variables names can be entered in text area under heading 'Enter Character names'. The character names should as small as possible and entered in exact sequence as your character appeared in data. Separate lines are used for each character name as shown the screenshot for help. After Entering the data and character names press **"Submit"** button.



Once the data is submitted for analysis, module will displaysnew web page asking to provide two information viz. Number of Characters and Number of observations per character as shown below in the screenshot



Fill information in the text boxes provided in front of each option. Press **"Analyse"** button.

The online module produces the output and displaysit on separate web page with following statistics:

Principal Component Analysis Based on Correlation Matrix

Table I. Descriptive Statistics

	Sepal Length	Sepal Width	Pedal Length	Pedal Width	Species
Mean	5.843	3.054	3.759	1.199	1.000
Variance	0.686	0.188	3.113	0.582	0.671
S.D	0.828	0.434	1.764	0.763	0.819

Covariance Matrix

	Sepal Length	Sepal Width	Pedal Length	Pedal Width	Species
Sepal Length	0.686	-0.039	1.274	0.517	0.531
Sepal Width	-0.039	0.188	-0.322	-0.118	-0.149
Pedal Length	1.274	-0.322	3.113	1.296	1.372
Pedal Width	0.517	-0.118	1.296	0.582	0.598
Species	0.531	-0.149	1.372	0.598	0.671

Correlation Matrix

	Sepal Length	Sepal Width	Pedal Length	Pedal Width	Species
Sepal Length		-0.109	0.872	0.818	0.783
Sepal Width	-0.109	1.000	-0.421	-0.357	-0.419
Pedal Length	0.872	-0.421	1.000	0.963	0.949
Pedal Width	0.818	-0.357	0.963	1.000	0.956
Species	0.783	-0.419	0.949	0.956	1.000

Eigen values of Correlation Matrix

	PC1	PC2	PC3	PC4	PC5
Eigenvalues	3.830	0.921	0.187	0.042	0.020
Proportion	0.766	0.184	0.037	0.008	0.004
Cumulative Proportion	0.766	0.950	0.987	0.996	1.000

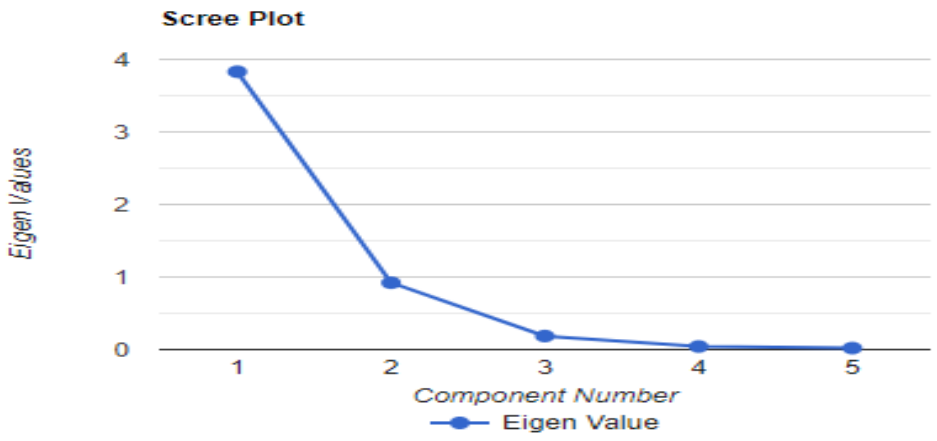


Table II.Loadings (Eigenvectors) of Correlation Matrix

	PC1	PC2	PC3	PC4	PC5
Sepallength	0.446	0.378	-0.752	0.141	-0.270
Sepalwidth	-0.229	0.923	0.285	0.005	0.122
Pedallength	0.507	0.026	-0.029	-0.247	0.825
Pedalwidth	0.497	0.070	0.387	-0.610	-0.476
species	0.495	-0.012	0.450	0.740	-0.067

Table III. Correlation of Principal Components with Original Variables

	PC1	PC2	PC3	PC4	PC5	Communality k=2
Sepallength	0.872	0.363	-0.325	0.029	-0.039	0.892
Sepalwidth	-0.447	0.886	0.123	0.001	0.017	0.985
Pedallength	0.991	0.025	-0.013	-0.051	0.118	0.983
Pedalwidth	0.973	0.067	0.167	-0.125	-0.068	0.952
species	0.969	-0.011	0.194	0.152	-0.010	0.939

Table IV. Principal Component Scores from Correlation Matrix

	PC1	PC2	PC3	PC4	PC5
O1	-1.312	0.518	-0.110	0.500	-0.215
O2	-1.232	-0.686	-0.452	0.306	-0.745
. . .	. . .	. . .	. . .	. . .	. . .
O148	0.982	0.277	0.694	0.843	-0.954
O149	0.921	1.053	2.278	-0.687	-0.135
O150	0.736	-0.029	1.725	1.192	0.960

IV. CONCLUSION:

To test the accuracy and reliability of the tool iris dataset was used and outputs were compared with standard statistical packages and it was found that this module produces the same output almost same degree of accuracy as that of standard packages. The scripting language ASP basedis convenient and easily available onlinetoolthat helps the stakeholders for performingprincipal component analysis.The tool is likely to be proved extremely helpful to researchers because of its capability to computedescriptive statistics such as mean, variance, standard deviation, Correlation/covariance matrices, eigenvalues, eigenvectors and component scores just in two steps.In addition to above this tool also computes the correlation for principal components with original variables and principal component scores.

The freely available online tool meets the growing needs of researcher to perform principal component analysis.It can save time by doing complex calculations automatically and generating results in understandable format .The tool is available online at http://14.139.232.166/pca/pca.html.

REFERENCES:

1. Abdi H, Valentin D and Edelman B. (1999). Neural Networks. Thousand Oaks, CA: Sage.
2. Diamantaras KI and Kung SY. (1996). Principal Component Neutral Networks: Theory and Applications. New York: John Wiley and Sons.
3. Dray S. (2008). On the number of principal components: a test of dimensionality based on measurements of similarity between matrices. Computational Statistics and Data AnalysisVol.(52), pp. 2228-2237.
4. Eastment HT and Krzanowski WJ. (1982). Cross-validatory choice of the number of components from a principal component analysis. Technometrics, Vol.(24), pp.73-77.
5. Escofier B and Pages` J. (1994). Multiple factor analysis, Computational Statistics and Data Analysis,Vol.(18), pp.121-140.
6. Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, Vol.(25), pp.417-441.
7. Jackson JE. (1991). A User's Guide to Principal Components. New York: John Wiley & Sons.
8. Jolliffe IT. (2002). Principal Component Analysis. New York: Springer.
9. Peres-Neto PR, Jackson DA and Somers KM. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited, Computational Statistics and Data Analysis, Vol.(49), pp. 974-997.

10. Sheoran, O.P; Tonk, D.S; Kaushik, L.S; Hasija, R.C and Pannu, R.S (1998). Statistical Software Package for Agricultural Research Workers, Recent Advances in information theory, Statistics and Computer Applications edited by D.S. Hooda and R.C. Hasija, Department of Mathematics and Statistics, CCSHAU,isar,pp.139-143.

AUTHORS PROFILE

O. P. Sheoran, Professor, Department of Mathematics and Statistics, College of Basic Sciences and Humanities, CCS Haryana Agricultural University, Hisar-125001

Vinay Kumar, Assistant Scientist, Department of Mathematics and Statistics, College of Basic Sciences and Humanities, CCS Haryana Agricultural University, Hisar-125001

Hemanat Poonia, Assistant Professor, Department of Mathematics and Statistics, College of Basic Sciences and Humanities, CCS Haryana Agricultural University, Hisar-125001

Komal Malik, Assistant Professor, Department of Economics, Government College, Nalwa, Hisar-125001