EXCELLENT PUBLISHERS

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 7 Number 05 (2018)**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2018.705.441

Principal Component Analysis Utilizing R and SAS Software's

Immad A. Shah[1*], Imran Khan[1], Shakeel A. Mir[1], M. S. Pukhta[1] and Ajaz A. Lone[2]

[1]Division of statistics, SKUAST-K, India
[2]Division of plant breeding, SKUAST-K, India

**Corresponding author*

ABSTRACT

In this study high end general statistical software's R and SAS have been compared using Principal component analysis. The Eigen values and Eigen vectors obtained by using R and SAS are found to be same, but the Eigen vectors obtained were found to differ in signs.

Keywords

R, SAS, Principal Component Analysis (PCA), Correlation, Eigen Values, Eigen Vectors

Article Info

Accepted:
30 April 2018
Available Online:
10 May 2018

Introduction

Statistical computing methods enable us to answer quantitative biological questions from research data and help plan new experiments in a way that amount of information generated from each experiment is maximized. Widespread use of high end statistical software packages have helped and greatly improved the ability of researchers to analyze and interpret voluminous data.

Developments in computerized statistical analysis have enhanced the ability of researchers to come up with better conclusions. This has helped in improving their statistical and computer related skills. For exploiting and sustaining these developed skills, high end general statistical software packages viz. R and SAS software have been used to perform a multivariate technique viz. PCA.

One of the biggest benefits of R is that it is completely free, and can be downloaded from www.r-project.org. A new version is released every six months, which means that any bugs are quickly fixed. R has a respected group of core developers who maintain and upgrade the basic R installation, but anyone can contribute add-on packages which provide additional functionality, such as specialized statistical tests or graphical functions. There are thousands of such packages available. R is not just a statistical package, but a programming environment, allowing users complete control over all aspects of data analysis. R can produce an enormous variety of production-quality graphical output in all of the standard formats. Modern research often involves collaboration between researchers with different scientific backgrounds and expertise. Knowing the basics of a tool that is commonly used by the bioinformatics, computational biology, and statistics communities will allow researchers to communicate better with the collaborators and to share information and data more easily.

The technique of Principal Component Analysis was first described by Karl Pearson (1901). Principal components are linear combinations of random or statistical variables which have special properties in terms of variance. A principal component analysis is concerned with the explaining of the variance-covariance structure through a few linear combinations of the original variables. It's seen as a technique for transforming a set of observed variables into a set of new variables that are uncorrelated with one another, Everitt (2005). Its general objectives are (1) data reduction (2) interpretation. Although *p* components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number, *k*, of the principal components. If so, there is almost as much information in the *k* components as there is in the original *p* variables. The *k* principal components can then replace the initial *p* variables and the original data set, consisting of *n* measurements on *p* variables is reduced to one consisting of *n* measurements on *k* principal components. Principal components depend solely on the covariance matrix Σ (or the correlation matrix ρ) of X_1, X_2...X_p. Their development doesn't require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be made from the sample components when the population is multivariate normal (Johnson amd Wichern, 1992).

Analysis of principal components is more of a means to an end rather than end in themselves because they frequently serve as intermediate steps in much larger investigations. A significant contribution towards the description of the practical computing methods for principal component analysis was due to Hotelling (1933). Principal components may be inputs to a multiple regression or cluster analysis. Moreover, principal components are one "factoring" of the covariance matrix for the factor analysis model. Modern competitors to principal component analysis that may offer more powerful analysis of the complex multivariate data are "projection pursuit" (Jones and Sibson, 1987), and "independent components analysis" (Hyvarinen *et al.,* 2001).

Materials and Methods

In the present study the data was obtained from DARS (Dryland Agriculture Research Station), SKUAST-Kashmir, comprising of 55 genotypes of maize. Twelve characters (Plant Height, Ear Height, Days to 50% Tasselling, Days to 50% Silking, 75% HB, Cob Length, Cob per Plant, Rows per Cob, Grains per Row, Cob Diameter, 100 Seed Weight, Yield per Plant) were evaluated for each genotype. R was downloaded from www.r-project.org. Venables and Repley (2004) have been used to have an understanding of this software. SAS 9.4 is a licensed platform and can be obtained from www.sas.com.

Results and Discussion

The basis for undergoing the multivariate analysis using principal component analysis is to check the correlation matrix whether the variables have some correlation or not. A high positive or negative correlation between the variables indicates that the variables are correlated and there is a sufficient reason to go for the Principal Component Analysis.

The correlation between the characters was obtained in the form of a correlation matrix and scatter plot using R software as shown in Table1 and Fig.1 respectively. The function for obtaining the Pearson's correlation is:

Pearson_cor <- cor(file)

Where "*Pearson_cor"* is the output name, *"cor"* is the command for obtaining the correlation matrix, and "*file*" is the data frame name. Several characters were found to be highly correlated such as grain/row and yield (r = 0.925), Days to 50% Tasseling and Days to 50% Silking (r = 0.939), Cob Length and Grain per Row (r = 0.942), 100 Seed Weight and Yield per Plant (r = 0.895), Cob Diameter and Cob Length (r = 0.8452), Row/Cob and Cob Diameter (r = 0.8594), Grain per row and Cob diameter (r = 0.8603), Cob Diameter and Yield per plant (r = 0.8816), Cob Length and 100 Seed Weight (r = 0.8130), Cob Length and Yield per Plant (r = 0.8827), Days to 50% Tasseling and Days to 50% Silking (r = 0.9387). Thus, there is a sufficient reason to go for principal component analysis.

For carrying out PCA in R software there are two functions. The generally preferred method for numerical accuracy is "*prcomp()*"where the calculation is done by a singular value decomposition of the centred and scaled data matrix, not by using eigenvalues on the covariance matrix (Rao,1964 and Jackson,1991) as in the alternative function "*princomp()*". The functions used to perform principal component analysis is *pca<- prcomp (file,center = TRUE,scale = TRUE).* Standardization ensures equal weightage of the measurements. Scaling of the data matrix can be done either in the *prcomp* function itself using arguments *"center=TRUE", "scale=TRUE"* or by executing the *"scale(file)"* command initially on the data matrix and the function *ev<-eigen(Pearson_cor)* is used to calculate the eigen values and eigen vectors in R software where *"ev"* is the output name. In SAS the *PROC FACTOR* function is used to obtain the principal components.

The results obtained are displayed in Table 2. It can be seen from Table 2 that the eigen values calculated from R and SAS yield same results upto three decimal places.

Eigenvalues represent the variances explained by the principal components (Table 2). The sum of the Eigenvalues (i.e. total variance), is equal to the trace of the diagonal elements of the correlation matrix. Highest eigenvalue was obtained for PC1 with an eigenvalue of 6.92 followed by PC2 with an eigenvalue of 2.30 indicating that the variance of PC1 is the largest of all, thus explaining maximum variability in the data. Since eigenvalue for other principal components is <1 hence are not retained.

The rule for choosing components with eigen value greater than 1 was originally suggested by Kaiser (1958). The proportion of variance explained by PC1 and PC2 are 57.69% and 19.23% respectively cumulating to 76.92% of the total variation. Eigenvector represent the coefficients of the principal components as shown in Table 3. It was found that eigen vectors calculated by using R and SAS give similar results upto four decimal places but opposite in signs. Here we see that different characters have different loadings on the eigenvectors. Most of the characters load on vector 1 and only four of the characters were found to load on vector 2. Loadings indicated by blank cells (--) are either zero or nearer to zero. For each component high loadings will result for few of the variables.

Table.1 Correlation Coefficients of characters of 55 Genotypes

Correlation Matrix												
	PlHt	**ErHt**	**Tsl**	**Sil**	**HB**	**CobLn**	**Cobpt**	**Rowcob**	**GrnRow**	**Cobdia**	**Sdwt**	**YPlnt**
PlHt	1.0000											
ErHt	0.8919	1.0000										
Tsl	-0.0819	-0.0679	1.0000									
Sil	-0.1444	-0.1106	0.9387	1.0000								
HB	0.1233	0.2235	0.4243	0.4692	1.0000							
CobLn	0.6497	0.7524	0.0114	0.0044	0.1494	1.0000						
Cobpt	0.4810	0.3674	-0.0644	-0.1300	-0.1174	0.3977	1.0000					
Rowcob	0.6287	0.7313	-0.0063	-0.0188	0.1895	0.7328	0.2976	1.0000				
GrnRow	0.7255	0.8123	0.0444	0.0350	0.2373	0.9422	0.4662	0.8021	1.0000			
Cobdia	0.6908	0.7631	-0.0770	-0.1022	0.1334	0.8452	0.4867	0.8594	0.8603	1.0000		
Sdwt	0.7819	0.8609	0.0401	0.0226	0.1846	0.8130	0.3490	0.7542	0.8587	0.7973	1.0000	
YPlnt	0.7473	0.7961	0.0079	-0.0249	0.0993	0.8827	0.5951	0.8141	0.9253	0.8816	0.8947	1.0000

*PlHt = PlantHeight, ErHt = Ear Height, Tsl = Days to 50% Tasseling, Sil = Days to 50% Silking, Hb = 75% HuskBrowning, CobLn = Cob Length, Cobpt = Cob per plant, Rowcob = Rows per cob, GrnRow = Grains per row, Cobdia = Cob Diameter, Sdwt = 100 Seed Weight, YPlnt = Yield per plant.

Table.2 Eigen Values obtained using R and SAS software

Principal Components		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Eigen Values	**R**	6.9225	2.3072	0.9014	0.6161	0.4961	0.3027	0.1545	0.1068	0.0769	0.0544	0.0338	0.0269
	SAS	6.9227	2.3074	0.9014	0.6165	0.4958	0.3027	0.1543	0.1067	0.0769	0.0547	0.0335	0.0270
ProportionVar		0.5769	0.1923	0.0751	0.0514	0.0413	0.0252	0.0129	0.0089	0.0064	0.0046	0.0028	0.0023
Cumm-Var		0.5769	0.7692	0.8443	0.8957	0.9370	0.9622	0.9751	0.9840	0.9904	0.9950	0.9977	1

Table.3 Eigen Vectors for PC1 and PC2 obtained using R and SAS

Characters		**Plant Height**	**Ear Height**	**50%Tasselling**	**50% Silking**	**75% HB**	**Cob Length**	**Cob/Plant**	**Row/Cob**	**Grains/ Row**	**Cob Diameter**	**100Seed Weight**	**Yield/Plant**
Eigen vector 1	**R**	-0.3198	-0.3425	0.0070	0.0188	-0.0721	-0.3442	-0.2004	-0.3259	-0.3622	-0.3504	-0.3496	-0.3654
	SAS	0.3198	0.3425	-0.0070	-0.0189	0.0721	0.3442	0.2004	0.3259	0.3622	0.3504	0.3496	0.3655
Eigen vector 2	**R**	-0.0670	0.0212	-0.6131	-0.6278	-0.4458	-0.0280	0.1331	-0.0264	-0.0575	0.0393	-0.0458	0.0053
	SAS	-0.0670	-0.0212	0.6131	0.6278	0.4458	0.0280	-0.1331	0.0264	0.0575	-0.0394	0.0458	-0.0053
Loadings using R	**Vector 1**	-0.320	-0.343	--	--	--	-0.344	-0.200	-0.326	-0.362	-0.350	-0.350	-0.365
	Vector 2	--	--	-0.613	-0.628	-0.446	--	0.133	--	--	--	--	--

Fig.1 Scatter Plot giving a graphical view of the correlations

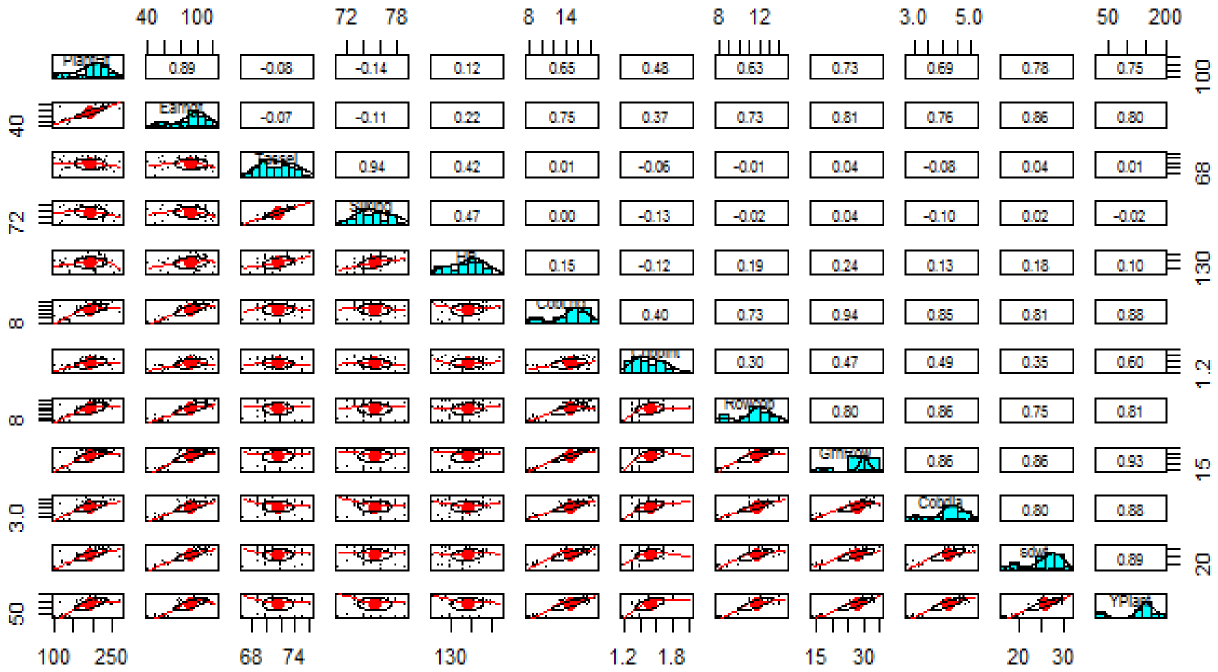
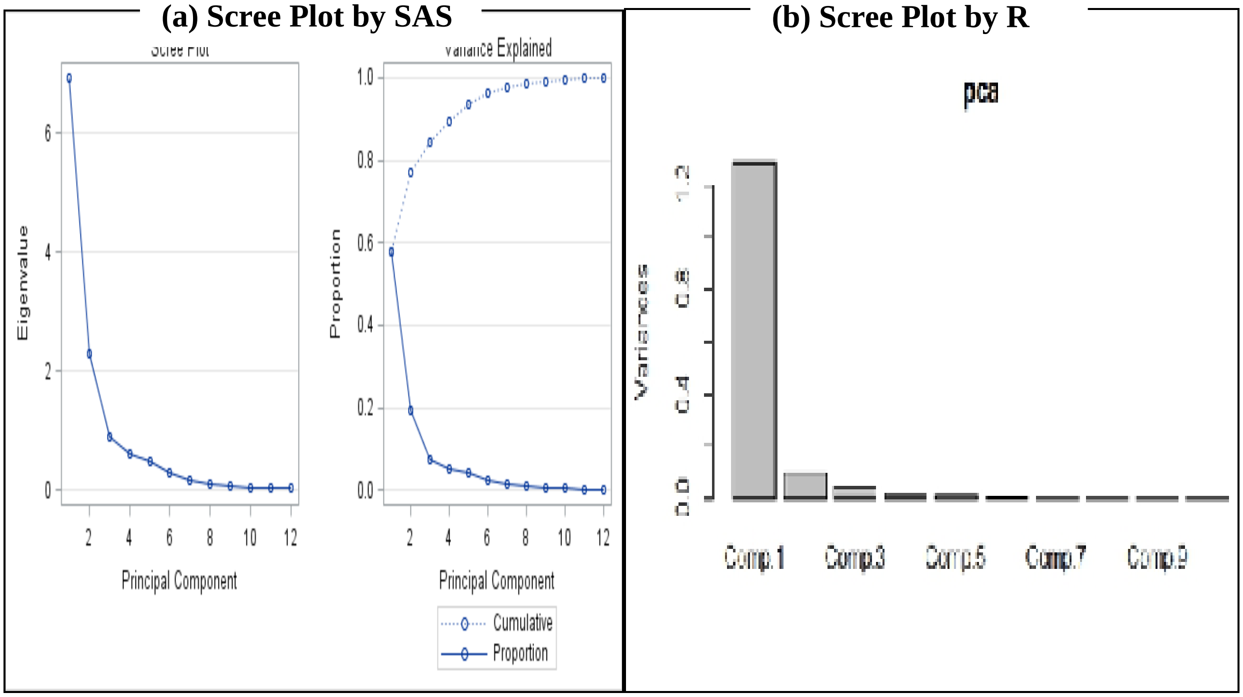
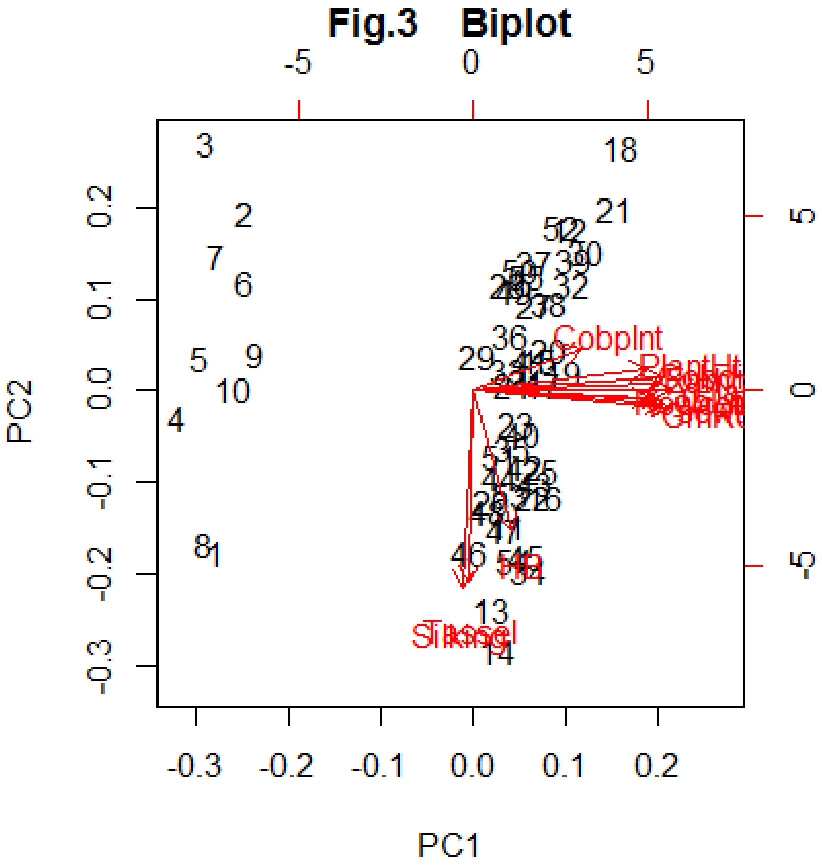


Fig.2 (a) and Fig.2 (b)


(a) Scree Plot by SAS

(b) Scree Plot by R

Fig.3 Biplot using R software relative loadings of the genotypes on the first and second principal components



To have a graphical view of the principal components, scree plot given by Catell (1966) for the eigenvalues of the principal components is obtained in R software using the *screeplot (pca)* function where *"pca"* is the name of the previously saved PCA using *"prcomp"* function as shown in Fig. 2(a). It can be seen from the Fig. 2(a) and Fig. 2(b) generated by using SAS and R that the rate of change stabilizes after PC2, giving an idea about the major principal components and number of principal components to select. In SAS software a separate function for generating scree plot Fig. 2(a) is not required in which the rate of change stabilises after second dot (.). The loadings obtained were plotted using a biplot graph. In a biplot, the original variables are shown by arrows (12 of them in this case) and can be obtained by using the *biplot (pca)* function in R, where *"pca"* is the name of the previously saved PCA as shown in Fig.3. The numbers represent the rows in the original data frame i.e. the genotypes, and the directions of the arrows show the relative loadings of the characters on the first and second principal components.

In vector 1, the important characters responsible for genetic divergence in the major axis of differentiation were Plant height, Ear height, Rows per cob, Cob per plant, Grain per row, 100 Seed weight, Cob length, Cob diameter and Yield per plant. In vector 2, which was the second axis of differentiation, the characters Days to 50% Tasseling, Days to 50% Silking, and 75% Husk Browning were found to be important. Same results were obtained by *PRINCOMP* procedure in SAS software.

The results obtained for both the R and SAS software's showed similar results but a difference in the signs of the eigenvectors was seen. The analysis of the data show eigen vectors with opposite signs i.e. the eigen vectors calculated by using SAS have positive sign, and have negative signs when calculated by using R software and vice versa. The signs of eigen vectors are arbitrary and does not affect the interpretation or the end result of the data. Since both R and SAS software showed only two eigen values >1 hence the 12 variables could be replaced by only two principal components PC1 and PC2. Also from scree plot it is seen that the rate of change is almost zero from PC2 onwards. The cumulative percentage proportion of variation explained by two PC1 and PC2 is 76.9%. Eigenvalues obtained from the R or SAS software represented the variances of the principal components with PC1 having an eigenvalue or variance of 6.92 and PC2 having an eigenvalue of 2.30. Similarly, the eigenvalues had associated eigenvectors which represented the coefficients or the loadings of the principal components. The characters were found to load on PC1 and PC2 separately. Most of the characters loading on PC1 indicate their relative importance. To ensure equal weightage to the variables measured on different scales, original variables were standardized. Working with R and SAS it was observed that the results displayed in R are function specific while in SAS a single function extracts a whole result for the statistical method employed. R because of its open source nature is a very cost-effective option and latest packages get released quickly. On the other hand SAS ends up as an expensive option.

References

Cattell, R. B. 1966. The screen test for the number of factors. *Multivariate Behavioral Research*, 1: 245-276.

Everitt, B. S. 2005. *An R and S-plus companion for multivariate analysis*. Springer.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational pshychology* 24: 417-41 and 498-520.

Hyvarinen *et al.*, 2001. *Independent Component Analysis*. Wiley, New York.

Jackson, J.E. 1991. *A User's Guide to Principal Components*. Wiley, New York.

Johnson, R.A. and Wichern, D.W. 1992. Applied Multivariate Statistical Analysis. 3rd Edition. Prentice-Hall International (UK) Limited, London

Jones, M.C. and Sibson, R. 1987. What is Projection Pursuit? *Journal of the Royal Statistical Society, Series A (General)* Vol. 150, 1: 1-37

Kaiser, HF. 1958. The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, 23:187-200

Pearson, K. 1901. On lines and planes of closest fit to a system of points in space. *Philosophical Magazine* 2: 557-72

R Core Team (2013). R: A language and enivironment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rao, C.R. 1964. The Use and Interpretation of Principal Component Analysis in Applied Research. Sankhyā: *The Indian*

Journal of Statistics, Series A (1961-2002) Vol. 26, 4: 329-358

SAS Institute Inc. 2011. Base SAS 9.3 Procedures Guide. Cary, NC: SAS Institute Inc.

Venables, W.N. and Ripley, B.D. 2009. Use of transformed auxillary variable in estimating the finite population mean. *Biometrical Journal* 41(5): 627-636

How to cite this article:

Immad A. Shah, Imran Khan, Shakeel A. Mir, M. S. Pukhta and Ajaz A. Lone. 2018. Principal Component Analysis Utilizing R and SAS Software's. *Int.J.Curr.Microbiol.App.Sci.* 7(05): 3794-3801. doi: https://doi.org/10.20546/ijcmas.2018.705.441