Advances in Bioresearch
Adv. Biores., Vol 10 (1) January 2019: 19-23

Print ISSN 0976-4585; Online ISSN 2277-1573
Journal's URL:http://www.soeagra.com/abr.html
CODEN: ABRDC3
DOI: 10.15515/abr.0976-4585.10.1.1923

Advances in Bioresearch

ORIGINAL ARTICLE

Exploratory Factor Analysis (EFA) Utilizing R Software Analytics

Immad A Shah[1], Imran Khan[2], Shakeel A Mir[3], M.S.Pukhta[4], Mohd Shafi Wani[5] Ajaz A Lone[6]
[1]Ph.D Scholar, Division of Agricultural Statistics, [2]Assistant Professor, Division of Agricultural Statistics, [3]Professor and Head, Division of Agricultural Statistics, [4]Associate Professor, Division of Agricultural Statistics, [5]Assistant Professor, Division of Agricultural Statistics [6]Assistant Professor, Division of Genetics and Plant Breeding,
Sher-e- Kashmir University of Agricultural Sciences and Technology, J&K, India 190025
Email: immad11w@gmail.com

ABSTRACT
In this study the data has been subjected to multivariate factor analysis utilizing R software using the Factanal function and Psych Package. The end results revealed that the variables under consideration can be grouped into two factors based upon the relative loading of each variable on a given factor. The rotation used is Varimax rotation, has as its rationale the aim of factors with a large few loadings and as many near zero loadings as possible. This is achieved by the iterative maximization of the quadratic function of the loading.
Keywords: *R Software, Factor analysis, Loadings, Communality, Uniqueness, Varimax*

Received 11.10.2018 Revised 28.11.2018 Accepted 19.12.2018

How to cite this article:
Immad A Shah, Imran Khan, Shakeel A Mir, M.S.Pukhta, Mohd Shafi Wani, Ajaz A Lone. Exploratory Factor Analysis (EFA) Utilizing R Software Analytics . Adv. Biores., Vol 10 [1] January 2019 19-23.

INTRODUCTION

Factor analysis can be considered as an extension of the principal component analysis and owes its development to Charles Spearman, 1904. Both can be viewed as attempts to approximate the covariance matrix Σ. However the approximation based on the factor analysis model is more elaborate. Factor analysis is specifically designed to look for meaningful commonality in a set of variables [1]. There are two types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA looks to explore the data to find an acceptable set of factors. CFA, on the other hand, begins with a theory or hypothesis about how the factors should be constructed and seeks to test whether the hypothesized structure adequately fits the observed data The primary question in factor analysis is whether the data is consistent with a prescribed structure. The observable random vector *X* with *p* components, has mean μ and covariance matrix Σ. The factor model postulates that *X* is linearly dependent upon a few unobservable random variables F_1, F_2,...,F_m called as "Common factors" and '*p*' additional sources of variations e_1, e_2, ..., e_p called errors or sometimes "specific factors". In particular the factor analysis model is:

$$X_1 = l_{11} F_1 + l_{12} F_2 + ... + l_{1m} F_m + e_1$$
$$X_2 = l_{21} F_1 + l_{22} F_2 + ... + l_{2m} F_m + e_2$$
.
.
.
.
$$X_p = l_{p1} F_1 + l_{p2} F_2 + ... + l_{pm} F_m + e_p$$

or in matrix notation;

$$X\,(p\times 1) = L(p\times m)\, F(m\times 1) + e(p\times 1)$$

The coefficient lij is called as the loading of the variable on the jth factor, so the matrix L is the matrix of factor loadings. By Factor rotation the solution is made more interpretable without changing its

underlying mathematical properties.

MATERIAL AND METHODS

The data for the study was obtained from the experimental trial maintained at DARS (Dryland Agriculture Research Station), SKUAST-Kashmir, and comprised of 55 genotypes of maize. Twelve characters viz. Plant Height, Ear Height, Days to 50% Tasselling, Days to 50% Silking, 75% HB, Cob Length, Cob per Plant, Rows per Cob, Grains per Cow, Cob Diameter, 100 Seed Weight, Yield per Plant) were evaluated for each genotype. The factor analysis model was fitted using the R statistical package [3], which is available at http://cran.r-project.org. In particular, we used the "factanal" function in the R Psych package to fit the factor analysis model and rotate the loadings to get the final solution [2].

RESULTS AND DISCUSSION

The basis for undergoing the multivariate analysis using factor 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 multivariate analysis, see Fig.1. 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:

lower=lowerCor(data)

"lowerCor" calls cor with use='pairwise', method='pearson' as default values and returns (invisibly) the full correlation matrix and displays the lower off diagonal matrix. Several characters were found to be highly correlated such as grain/row and yield (0.93), 50% Tasseling and 50% silking (0.94), cob length and grain per row (0.94), and 100 seed weight and Yield (0.89). Thus, there is a sufficient reason to go for factor analysis.

Using the "*pairs.panels*" function to graphically show relationships. The x axis in each scatter plot represents the column variable, the y axis the row variable. The plot character was set to a period (pch='.') in order to make a cleaner graph. The command used is :

pp=pairs.panels(data,pch='.')

where *pp* is output name and *data* is the data frame to which the factor analysis is being done.

The heat map display of the correlational structure was obtained which gave perhaps a better way to see the structure in a correlation matrix as shown in figure 2. This is just a matrix color coded to represent the magnitude of the correlation and is useful when considering the number of factors in a data set. The color coding represents a "heat map" of the correlations, with darker shades of red representing stronger negative and darker shades of blue stronger positive correlations. The function "*cor.plot*" of the Psych package is used to obtain the heat map display. The command used is :

Heat = cor.plot(data,numbers=TRUE)

Where Heat is the output file name & data is the name of the data frame to which the analysis is being done.

The heat map shows a clear 4 factor solution thus giving an idea of how many factors to consider. The cor.plot function to show the correlations in a circumplex. Correlations are highest near the diagonal, diminish to zero further from the diagonal, and the increase again towards the corners of the matrix. Circumplex structures are common in the study of affect.R has multiple functions that will do factor extraction. As part of R's native packages, the *factanal* function will do maximum likelihood extraction. The command for doing factor analysis in R software is:

EXPLORATORY FACTOR ANALYSIS (EFA)….

> factor=factanal(data,factors=2,rotation="varimax")
> factor=factanal(data,factors=3,rotation="varimax")
> factor=factanal(data,factors=4,rotation="varimax")

Here the output variable has been named as factor. In the command "*data*" is the name of the data set to which factor analysis is to be done, next to that is the number of factors we want to obtain and we have set that equal to 4 as indicated by the heat map of the correlation structure. Rotation "*varimax*" is an orthogonal rotation which means that the new factors derived from the provisional factors will be uncorrelated and works upon the iterative maximization of a quadratic function of the loadings, Mardia *et al* [6]. As defined, the varimax procedure finds an orthogonal transformation matrix. Kline suggests that the most accepted method for creating factors with simple structure is varimax [5]. The results from the test of 4 factors suggest that a 2 factor model is adequate for the data. Hence we opt the two factor model.

The results obtained using the R software is given in the Table 2.

Uniqueness obtained by *Factanal* function for 2 factor model is shown in Table 3. Uniqueness is the part

of the variance associated with the error term. It is also known as specificity, the part of the variance that is unrelated to the common factors. The uniqueness or specificity being high for cob per plant. The *factanal* function also gives the factor loadings. Each factor has a factor loading associated with it specific to all the characters. The factor loadings for the factors 1 and 2 are given in Table 4.

The above factor analysis indicates that most of the variance for variables plant height to yield per plant is accounted for by the two factors. Table 5 shows the communalities associated with each of the characters. It is defined as the sum of the square of the factor loadings. The sum of the communality and the corresponding uniqueness should always be equal to one

From Table 4 it is observed that Plant Height, Ear Height, 75% HB, Cob Length, Cob per Plant, Row per Cob, Cob Diameter, Grain per Row, 100 Seed Weight and Yield per Plant are found to load almost entirely on Factor 1 with high loadings for plant height, ear height, cob length, row per cob, cob diameter, grain per row, 100 seed weight and yield per plant. While as 50 % Tasselling, 50 % Silking and 75% Husk Browning variables were found to load on Factor 2 with high loadings for 50 % Tasselling and 50% Silking. Factor scores or "factor loadings" indicate how each "hidden" factor is associated with the "observable" variables used in the analysis. To obtain the factor scores the following command is used in R,shown in Table 6 :

>fact_scores=factanal(data,factor=2,rotation="varimax",scores="regression")
>fact_scores$scores
>head(fact scores)

	Plant Ht	Ear hgt	50% T	50% S	75% HB	CobLng	Cob/plnt	Row/cob	Grn/Row	Cob dia	100 sd wt	Y/Plant
Plant Ht	1.00											
Ear hgt	0.89	1.00										
50% T	-0.08	-0.07	1.00									
50% S	-0.14	-0.11	0.94	1.00								
75% HB	0.12	0.22	0.42	0.47	1.00							
Cob Lng	0.65	0.75	0.01	0.00	0.15	1.00						
Cob/plnt	0.48	0.37	-0.06	-0.13	-0.12	0.40	1.00					
Row/cob	0.63	0.73	-0.01	-0.02	0.19	0.73	0.30	1.00				
Grn/Row	0.73	0.81	0.04	0.04	0.24	0.94	0.47	0.80	1.00			
Cob dia	0.69	0.76	-0.08	-0.10	0.13	0.85	0.49	0.86	0.86	1.00		
100 sd wt	0.78	0.86	0.04	0.02	0.19	0.81	0.35	0.75	0.86	0.80	1.00	
Y/Plant	0.75	0.80	0.01	-0.03	0.10	0.88	0.60	0.81	0.93	0.88	0.90	1.00

Table 1: Correlation between the characters as a matrix.

	2 Factor model		3 Factor model			4 Factor model			
	Factor1	Factor2	Factor1	Factor2	Factor3	Factor1	Factor2	Factor3	Factor4
SS Loading	6.697	2.162	5.622	2.159	1.631	5.618	2.151	1.457	0.586
Prop. Variance	0.558	0.180	0.468	0.180	0.136	0.468	0.179	0.121	0.049
Cummulative Variance	0.558	0.738	0.468	0.648	0.784	0.468	0.670	0.789	0.818

Table 2: Results for the formal test of the number of the factors.

Characters	Uniqueness
Plant Height	0.370
Ear Height	0.257
50% Taselling	0.114
50% Silking	0.005
75% HB	0.744
Cob Length	0.142
Cob/Plant	0.734
Row/Cob	0.293
Grains/Row	0.066
Cob Diameter	0.169
100 Seed Weight	0.173
Yield/Plant	0.073

Character	Factor 1	Factor 2
Plant Height	0.791	-
Ear Height	0.862	-
50% Taselling	-	0.938
50% Silking	-0.100	0.992
75% HB	0.140	0.486
Cob Length	0.921	-
Cob/Plant	0.509	-
Row/Cob	0.838	-
Grains/Row	0.957	0.132
Cob Diameter	0.912	-
100 Seed Weight	0.902	0.114
Yield/Plant	0.960	-

Table 3 & Table 4: Uniqueness obtained by *factanal* function & Factor loadings of the characters respectively.

Characters	Uniqueness	Communality
Plant height	0.370	0.6256
Ear height	0.257	0.7430
50% teaselling	0.114	0.8798
50% silking	0.005	0.9940
75% HB	0.744	0.2557
Cob length	0.142	0.8537
Cob/plant	0.734	0.2590
Row/cob	0.293	0.7022
Grains/row	0.066	0.9332
Cob diameter	0.169	0.8317
100 seed weight	0.173	0.8266
Yield/plant	0.073	0.9216

Table 5: Uniqueness and communality obtained from *factanal* function

Factor 1	Factor 2
-2.1419	1.6001
-1.6684	-1.8179
-1.9033	-2.3605
-2.2368	0.1100
-2.2326	0.1028
-1.7985	-0.8519

Table 6 : Factor Scores obtained from *factanal* function

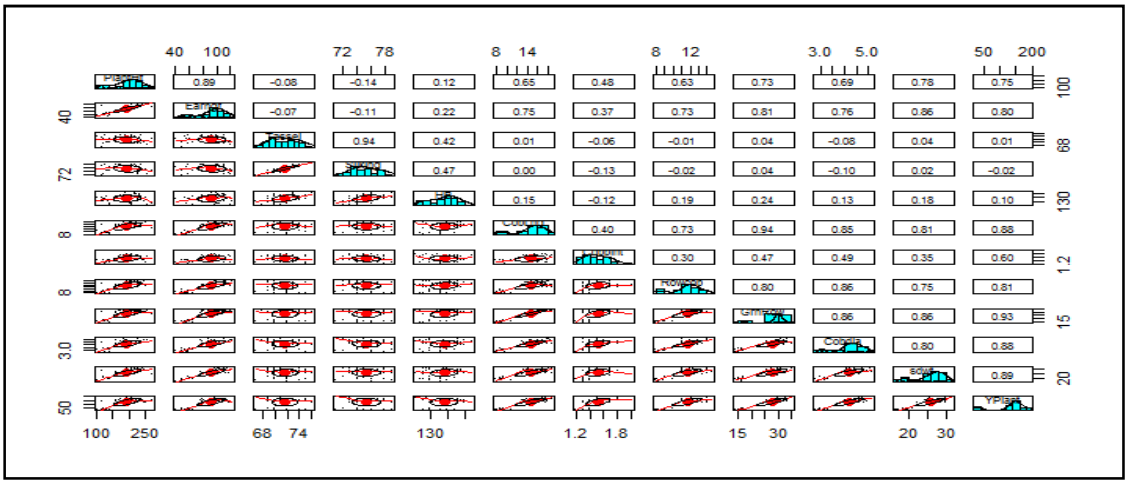

Figure 1: Scatter plot obtained from the pairs panel function.

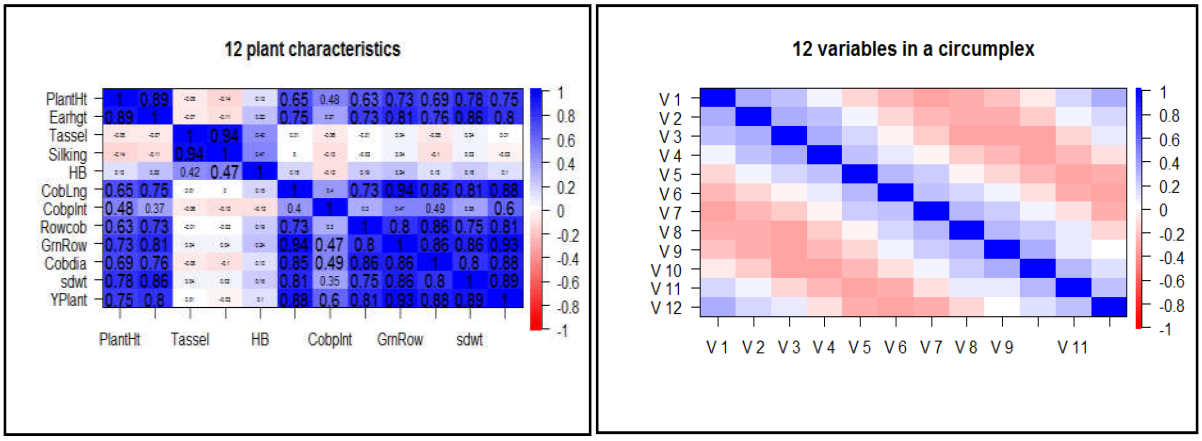

Figure 2: Heat Map representing the correlation between the characters.

CONCLUSION

The data when multivariate analysed using factor analysis suggested a 2 factor solution. The Heat map and the usual factanal function suggested a 4 factor solution. But keeping in view the cumulative variance of the number of factors, a 2 factor model explains 73.8% of the total variability in the data and there is not a considerable change in its cumulative variance as we move on to a 3 or 4 factor model. Hence a 2 factor model was adopted. The 1st factor involves characters Plant Height, Ear Height, 75% HB, Cob Length, Cob per Plant, Row per Cob, Cob Diameter, Grain per Row, 100 Seed Weight and Yield per Plant and we may label it as a factor of morphological traits. The 2nd factor comprises of characters 50 % Tasselling, 50 % Silking and 75% Husk Browning. The obvious label for the factor is reproductive traits.

REFERENCES

1. DeCoster, J. (1998). Overview of Factor Analysis. Accessed May 5, 2012 at www.stat-help.com/notes.html.
2. Revelle, W. (2011). Psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.0-98.
3. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
4. Kline, P. (1994). An Easy Guide to Factor Analysis. Routledge.
5. Mardia, K.V, Kent, J.T. and Bibby, J.M. (1979). Multivariate Analysis. London: Academic Press.