Journal of Pharmacognosy and Phytochemistry

Available online at www.phytojournal.com

J P P Journal of Pharmacognosy and Phytochemistry

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2018; SP1: 76-79

Hemant Kumar Singh
Mahatama Gandhi Inter College, Sakhawaniya, Kushinagar, Uttar Pradesh, India

Ravi Shankar Maurya
Department of Mathematics and Statistics, I.I.T., Kanpur, Uttar Pradesh, India

Bhim Singh
Department of Basic Science, College of Horticulture and Forestry, Jhalrapatan, Jhalawar, Rajasthan, India

Correspondence
Bhim Singh
Department of Basic Science, College of Horticulture and Forestry, Jhalrapatan, Jhalawar, Rajasthan, India

National Conference on Conservation Agriculture
(ITM University, Gwalior on 22-23 February, 2018)

Application of PCA in evaluation of wheat varieties

Hemant Kumar Singh, Ravi Shankar Maurya and Bhim Singh

Abstract
In the present paper, principal components analysis technique has been described to ascertain the best wheat verity by taking proper weights of important characters in the plant breeding experiments. The proposed method has been done for both using correlation matrix and variance covariance matrix. For illustration purpose, we used the data based on ten wheat varieties of multiple characters with three replications in randomized block design.

Keywords: Correlation coefficient, Correlation matrix Weight, Principal component analysis (PCA), Randomized block design (RBD)

Introduction
Principal component analysis, abbreviated as PCA, is a method of statistical analysis useful in data reduction and interpretation of multivariate data sets (Jackson, 1991) [7]. The consequence of the study for PCA was recognition as a statistical method essential in analyzing large data sets. PCA has now spread to a large number of scientific fields and is used in a variety of different applications where analysis of inter-object correlations is the focus. In particular, the past decades have seen growing use of PCA in cosmology as the field confronts rising statistical challenges.

A general statement of the problem solved by PCA is the following: Analyze the relationship between the m parameters of n objects provided a given m×n data set. The central step of PCA is the redefinition of this data set in terms of a new set of variables which are mutually-orthogonal linear combinations of the original variables. The new variables define a coordinate axis in multivariate data space that forms a ‘natural’ classification of the data. The motivation behind this redefinition is the dimensionality reduction achieved by an orthogonal basis. PCA condenses correlations in the data to single variables finding the ‘true dimensionality’ of the data set. Provided a willingness to sacrifice some accuracy for economy of description, there is the huge potential for significant reduction in data dimensionality. This makes PCA a powerful and versatile analysis method.

In statistical practice, the method of principal component is used to find the linear combinations with large variance. In many exploratory studies, the number of variables under consideration is too large to handle. Since it is the deviations in these studies that are of interest a way of reducing the number of variables to be treated is to discard, the linear combinations which have small variances and study only those with large variance. The principal components give a new set of linearly combined measurements. It may be that most of the variation from individual to individual resides in three linear combinations. Hotelling (1933) [5] indicated that principal component analysis (PCA) is an exploratory tool designed by Pearson (1901) [9] to identify unknown trends in a multi-dimensional data set. Therefore, it becomes impossible to make a visual inspection of the relationship between genes or conditions in such a multi-dimensional matrix. One way to make meaning of this data is to reduce its dimensionality (Hotelling, 1933). Several data decomposition techniques are available for this purpose and multivariate data analysis (Cooley and Lohnes, 1971) [3]. PCA is among these techniques that reduced the data into two dimensions (Smith, 2002; Rao, 1964; Raychaudhuri *et al.*, 2000) [13, 10, 11]. Multivariate analysis has been used frequently for genetic diversity analysis in many crops such as barley (Cross, 1992) [4], Sorghum (Ayana and Bekele, 1999) [1], peanut (Upadhyaya *et al.*, 2009) [14] and vineyard peach (Nikolic *et al.*, 2010) [8], rice (Bharadwaj *et al.*, 2001) [2] and in wheat (Hailu *et al.*, 2006 and Singh and Singh, 2013) [5, 12].

In most of the varietal trial in plant breeding experiments, the plant breeders generally study the characteristics of the varieties separately and make a trade-off among the characteristics to arrive most suitable variety. Among the characteristics of a variety the most economic trait is yield potential of a variety which may be of at most importance for an agronomist but plant breeder has to select a variety based on various characteristics for the selection process in the breeding programme. Generally, the plant breeding utilize the selection index based on multiple characteristics of the variety. However, the method of principle component analysis can suitably be used to study the relative importance of varieties based on multiple characteristics. Keeping in view of this fact, the objective of present investigation is to proposed indices for comparison of the varieties when certain weights are attached.

Materials and Methods

In agriculture, several types of experiments are conducted. Sometimes it is the varietal trial when the different varieties are tried in field in some of the situations. The experiments are also done by varieties in combination with several other variables called factors. Usually our problem is to ascertain the best variety out of several varieties tried in the field experiments. Beside this, in agricultural field experiments multiple characters are observed. For example, if there are number of variables affecting the crop of the variety, we do not make any consideration of all the factors at a time. Considering all the variables affecting the crop of the variety does not draw statistical inference. The variables are studied separately and for different factors separate conclusions are drawn. In the present investigation, our objective is to study the affect of several factors on the response of the variety simultaneously; no separate analysis of the character is made. Statistical inference is done by taking into account all the factors at a time.

The data are first analyzed for each of the characters/variables. The variance-covariance matrices of the variables are also worked out. Using variance-covariance matrix, the correlation coefficients between all pairs of variables are obtained. Next, the linear function L; corresponding to maximum eigen value of the correlation matrix is obtained. Now if the experiment is a varietal trial then for each variety, say, the ith one the mean $\overline{Z}_{ij}$ is obtained from

$$\overline{Z}_{ij} = \frac{\overline{X}_{ij} - \overline{X}_i}{S_i / \sqrt{r}}$$

where $\overline{X}_{ij}$ is the ith character mean for the jth variety and $\overline{X}_i$ is the overall mean for ith character and S_i^2 is the pooled variance of the ith character over varieties and r is the number of replications of the experiment. These values of Z_{ij} 's are substituted in L giving.

$$L_{1(j)} = \sum_{i=1}^{k} l_i \overline{Z}_{ij} \ , \quad j = 1, 2, ..., v$$

We shall be getting v values of $L_{1(j)}$. Infact, $L_{1(j)}$ is the first principal component with respect to jth varieties. These first principal components for different varieties are considered as indices for these varieties. The comparison of these $L_{1(j)}$ will reveal the relative position of the varieties. Only first principal component has been used to construct the indices for the varieties as it has maximum variability (information). Although several indices can be constructed using 2nd, 3rd principal components and so on, but only first principal component has been utilized. These indices are obtained purely from consideration of variance and covariance over the characters. This approach, however, does not seem to be fully realistic because while assessing the performance of a variety taking into account a number of characters of each of the varieties it is necessary to take into account the relative preference of the different characters. For example, if there is a character which provides a larger weight in the linear function from considerations of varieties and covariance's among the characters but it is not so much important in deciding the worth of a variety then the result obtainable from a weighted component may not be realistic from considerations of worth of the varieties. Hence, proper weights have to be given to the characters while getting the components. One way is to write the components as below:

$$L_j = \sum_{i=1}^{k} l_i w \overline{Z}_{ij} \ ; \ \sum w_i = 1 \, ,$$

We have used correlation matrix for finding indices. As the correlation coefficients are independent of any weighting factor of individual variables, the above method cannot be applied where the characters have to be weighted differently which is often necessary as it makes the investigation more realistic.

Accordingly, we discuss below a modified procedure which takes into account the different weights of different characters.

Let $X_1, X_2, \ldots, X_k$ be k different characters observed from an agricultural or other similar experiment. Let X be a matrix of data of order $n \times k$ on k variables and each variable has n observation. We also assume that X_i, ($i = 1, 2, ..., k$) has mean zero. We have to draw inference taking into account all these variables after attaching appropriate weights to the characters. These weights are given and are decided by the subject matter specialists. Let the weights of the characters be $W_1, W_2, ..., W_k$ respectively.

We now define $Y_i = W_i X_i$, $i = 1, 2, ..., k$. Next the values of the variable y_i are analysed for the experiment for which the variables X_i are observed. Actually, the results of analysis of variance of the Y_i variate can be obtained from those of the corresponding X_i variables simply by multiplying each sum of squares by W_i^2, $i = 1, 2, ..., k$. However, choice of W_i is the difficult in practice. To overcome this problem W_i is chosen as imaginary value in order of relative importance of the varieties based on their yield. Therefore, rows of X will interchanged accordingly as a result, the correlation matrix will also be changed. We want a linear function of the Y_i 's which has the maximum variance.

Let

$$L = \sum_{i=1}^{k} l_i y_i ,$$

where l_i are unknown constants to be determined.

We further take the restriction $\sum_{i=1}^{k} l_i^2 = 1$ and maximise $Var\ (L)$ subject to this restriction

$$Var\ (L) = \sum_i l_i^2 S_i^2 + \sum_i \sum_j l_i l_j Cov\ (Y_i, Y_j)$$

$$= \sum_i l_i^2 S_i^2 + \sum_{i<j} \sum_j l_i l_j S_i S_j r_{ij}$$

where S_i^2 is the variance of Y_i and r_{ij} is the correlation between ith and jth characters.

To find l_i's we have to maximize

$$E = \sum l_i^2 S_i^2 + \sum \sum l_i l_j S_i S_j r_{ij} - \lambda \left(\sum l_i^2 - 1 \right)$$

Differentiating E partially w.r.t. to l_i and equating to zero we get the following equations:

$$l_1 (S_1^2 - \lambda) + l_2 S_1 S_2 r_{12} + \ldots + l_k S_1 S_k r_{1k} = 0$$
$$l_1 S_1 S_2 r_{21} + l_2 (S_2^2 - \lambda) + \ldots + l_k S_2 S_k r_{2k} = 0$$
$$\ldots \quad \ldots \quad \ldots \quad \ldots$$
$$l_1 S_1 S_k r_{k1} + l_2 S_2 S_k r_{k2} + \ldots + l_k (S_k^2 - \lambda) = 0 \quad \text{(A)}$$

These equations are homogeneous. Non-trivial solution of these can be obtained only if these equations are dependent, that is the determinant formed of the coefficients l_i's of equations is zero. This determinant contains the Lagrange multiplier λ which is an unknown. Thus, equating the determinant to zero we get a polynomial equation in λ. Its degree is the same as the order of the determinant, that is k. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the solutions of these equations with λ_1 as the highest value among solutions. Now substituting λ_i for λ in equation at A, we shall get non- trivial solution of l_i's. These equations are the eigen vectors corresponding to Eigen value λ_i. Let these solutions be denoted by $\hat{l}_i$, $(i = 1, 2, \ldots, k)$.

Next, the function $\sum \hat{l}_i y_i$ are used as an index for drawing inference as required. The objectives of the experiment determine nature of inference required. If the experiment is a varietal trial we take X_i as the ith character mean for, say, the jth variety minus the overall mean of the ith character over all the varieties and then denote it by $\overline{Z}_{ij}$. Next, we use $L_{1(j)} = \sum l_i \overline{Z}_{ij}$ as the index. These results are discussed with the help of empirical study.

The present investigation is based on the experimental data taken from Crop Research Station, G.B. Pant University of Agriculture and Technology, Pantangar, Uttarakhand on ten varieties. The experiment was conducted with three replications and ten wheat varieties in a randomized black design. The data were obtained on the following characteristics.

X_1 = Days to flowering, X_2 = Days to maturity, X_3 = Plant height (cm.), X_4 = No. of tillers/plant, X_5 = Spike length (cm.), X_6 = Grains/spike, X_7 = No. of Spiklet/spike, X_8 = Grain weight/spike (gm.), X_9 = 100 grain weight, X_{10} = Yield/plant (gm.).

The experimental data as above mentioned character were obtained and means of the characters of ten varieties (RAJ 3777, Tepoka, HD 2285, Sonalika, HD 2278, HD 2444, CPAN 1666, HD 2270, HW 517 and UP 115) are given in Appendix I. The data were analyzed using the method described in previous section and are presented in the subsequent section.

Results and Discussion

For the illustrations of the methodology, a set of experimental data were used. The results are described in this section. The correlation matrices have been computed using means of the characteristics (Appendix I) and is presented in Table 1. The rows of the correlation matrix are interchanged which are considered as imaginary weights of the characters. The imaginary weights are attached to the characters on the basis of yield of the varieties. This resulted interchanging of the rows of original data set. Using the new matrix of correlation coefficients, the usual method of principal component is applied. Here first Eigen value is considered for computations. From Table 2, the first three preferred varieties in order were found to be RAJ 3777, Tepoka and Sonalika. However, this method provides substantial weight to the first preferred varieties and is true for other varieties too in subsequent positions.

Table 1: Correlation matrix of the characters

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*X8*	*X9*	*X10*
X1	1									
X2	-0.0705	1								
X3	-0.3134	0.6243	1							
X4	0.1641	-0.6086	-0.4116	1						
X5	-0.2889	0.6847	0.7578	-0.4352	1					
X6	-0.1735	0.7051	0.6892	-0.5007	0.5857	1				
X7	-0.1255	0.7311	0.6753	-0.7950	0.5872	0.8732	1			
X8	-0.4847	0.7308	0.5214	-0.5440	0.5553	0.8584	0.7349	1		
X9	-0.8011	-0.0288	0.0174	0.1532	0.0865	-0.1447	-0.2415	0.2050	1	
X10	-0.6345	0.2495	0.3724	-0.0639	0.0428	0.1998	0.0819	0.4451	0.4062	1

Table 2: Value of Eigen value, Eigen vector and varieties indices using usual principal component method with suitable weights.

B. No.	Name of variety	Eigen vector	Varietal indices	Relative Position of variety based on indices (Rank)
l.	RAJ 3777	0.3094	13.29	1
2	Tepoka	-0.8824	-23.38	2
3	HD 2285	-0.813	-62.42	4
4	Sonalika	0.6505	-32.77	3
5	HD2278	-0.7712	-70.27	5
6	HD 2444	-0.8993	-112.27	6
7	CPAN 1666	-0.9066	-154.77	7
8	HD 2270	-0.8748	-196.18	9
9	HW517	0.093	-191.81	8
10	UP 115	-0.319	-211.15	10

Eigen value: 5.0603

Conclusion

In the present Investigation, the inference is to be drawn by associating appropriate weights to the characters/indices based on principal components have been worked out for comparison among the varieties. The scores of varieties are assigned of the relative importance of individual varieties. The method has been explained for randomized block design (RBD) taking correlation matrix as a weight. The data collected refers to ten varieties and ten characters with three replications in randomized block design. In the present study, the method of principal component analysis has been used by considering interchanged correlation matrix as weights of the varieties. We observed on the basis of our analysis is that from the experimental data, wheat variety RAJ 3777 was found to be most preferred among all the ten varieties.

References

1. Ayana A, Bekele E. Multivariate analysis of morphological variation in sorghum (*Sorghum bicolor* L. Moench) germplasm from Ethiopia and Eritrea. Genet. Resource. Crop Evol. 1999; 46:273-284.
2. Bharadwaj Ch, Tara SC, Subramanyam D. Evaluation of different classifactory analysis methods in some rice (*Oryza sativa* L.) collections. Ind. J Agric. Sci. 2001; 71(2):123-125.
3. Cooley WW, Lohnes PR. Multivariate Data Analysis. John Wiley & Sons, Inc., New York, 1971.
4. Cross RJ. A proposed revision of the IBPGR barley descriptor list. Theor. Appl. Genet. 1992; 84:501-507.
5. Hailu F, Merker A, Singh H, Belay G, Johansson E. Multivariate analysis of diversity of tetraploid wheat germplasm from Ethiopia. Genet. Resource. Crop Evol. 2006; 54:83-97.
6. Hotelling H. Analysis of a complex of statistical variable into principal components. J Educ. Psych. 1933; 24:417-441.
7. Jackson JE. A User’s Guide to Principal Components. New York: Wiley-Interscience, 1991.
8. Nikolic D, Rakonjac V, Milatovic D, Fotiric M. Multivariate analysis of vineyard peach (*Prunus persica* L. Batsch.) germplasm collection. Euphytica. 2010; 171:227-234.
9. Pearson K. On lines and planes of closest fit to systems of points in space. Phill Mag 1901; 6(2):559-572.
10. Rao CR. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya* A. 1964; 26:329-358.
11. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: Application to sporulation time series. Pacific Symposium on Bio-computing, 2000.
12. Singh HK, Singh B. A note on multivariate technique in agricultural experiments. International Journal of Current Science. 2013; 5:50-56.
13. Smith LI. A tutorial on Principal Components Analysis, 2002. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.
14. Upadhyaya HD, Reddy LJ, Dwivedi SL, Gowda CLL, Singh S. Phenotypic diversity in cold-tolerant peanut (*Arachis hypogaea* L.) germplasm. Euphytica. 2009; 165:279-291.