Journal of Applied Sciences 5 (1): 113-117, 2005
ISSN 1607-8926


# Multivariate Analysis of Mash Data

[1]I.A. Arshad, [1]F. Muhammad and [2]A. Ghafoor
[1]Department of Mathematics and Statistics, Allama Iqbal Open University, Islamabad, Pakistan
[2]Department of Statistics, Government College Farooka District, Sargodha, Pakistan

**Abstract:** Different mash plant traits contributes to the mash grain yield but the major contributors are plant height ($X_1$), days to flowering ($X_2$), days to first pod maturity ($X_3$), days to 90% maturity ($X_4$), branches per plant ($X_5$), pods per plant ($X_6$), pod length ($X_7$), seeds per pod ($X_8$), 100-seed weight ($X_9$), biological yield per plant ($X_{10}$) and mash grain yield (Y). This study was initiated to find the important regressors on which the yield of mash depends. In this regard principal component analysis and path analysis were used to find correlation structure between mash plant traits and regressors effect on mash grain yield, respectively. Principal component analysis reduced the dimensionality in the system of eleven mash plant traits to four principal components, which contributes about 88% of the total variability present in the mash plant data. On the basis of correlation between principal components and original mash plant traits, a classification structure was made to observe the relation between different traits. It was observed that for the first principal component, plant height ($X_1$), days to flowering days to first pod maturity ($X_3$), days to 90% maturity ($X_4$) and 100 seeds weight ($X_9$) have positive correlation between themselves i.e. varies in the same direction. Path analysis was also described to explain correlation structure, direct-indirect effects between different mash plant traits. This analysis suggested that pod per plant has maximum positive direct effect on mash grain yield i.e. more pod per plant, greater will be the yield. But days to 90% maturity has maximum negative direct effect on mash grain yield i.e. more maturity lesser will be the grain yield. Similarly branches per plant and biological yield per plant have positive indirect effect on mash grain yield via pods per plant. It was observed that the direct and indirect effects of remaining predictors are negligible.

**Key words:** Principal components, path analysis, scree plot, residual

## INTRODUCTION

Achievement of self-sufficiency in agricultural production is very important step for economic growth of any country. In the present study, several results on Mashbean (*Vagna mungo*) were evaluated, which is an important Kharif crop of Pakistan and a rich source of quality proteins. The breeders are usually interested to develop the genotypes, which are high in yield and separation of mash plant traits which are contributing positively and negatively towards mash grain yield and check their direct and indirect influence on the mash grain yield.

Shukla[1] analyzed genotype x interaction in a series of winter wheat variety trials over three years by various techniques. He regressed genotypic values on the environmental mean and calculated principal components of variation within genotypes. The conclusion from this set of analysis is the interaction contained three or four significant principal components.

Ayub *et al.*[2] applied correlation and path coefficients analysis on brassica genotypes data. Path coefficient analysis revealed that branches per plant had maximum direct effect on seed yield. 1000-seeds weight has a negative indirect effect on seed yield via branches per plant and pod length.

Kubsad *et al.*[3] estimated correlation and path coefficients for 9 traits of safflower and showed that the highest association was revealed between seed yield and 100-seed weight. Path coefficient analysis showed that the dry matter per plant had maximum contribution towards seed yield and was closely followed by seeds per main capitulum.

## MATERIALS AND METHODS

Mash data was obtained form Plant Genetic Resource Institute at National Agricultural Research Center Islamabad. The experimental material consists of 37 mash genotypes arranged in Randomized Complete Block

**Corresponding Author:** Dr. Irshad Ahmed Arshad, Department of Mathematics and Statistics, Allama Iqbal Open University, Sector H-8, Islamabad, Pakistan Tel: 051 9257061 E-mail: irshad_arshad@hotmail.com

Design with three replications. Eleven different characters including morphological characters were measured such as plant height ($X_1$), days to flowering ($X_2$), days to first pod maturity ($X_3$), days to 90% maturity ($X_4$), branches per plant ($X_5$), pods per plant ($X_6$), pod length ($X_7$), seeds per pod ($X_8$), 100-seed weight ($X_9$), biological yield per plant ($X_{10}$) and mash grain yield per plant (Y). Principal component analysis was used to explain the correlation structure of yield related characters of mashbean. Path analysis was used to measure both direct and indirect effects that morphological character may have upon the grain yield of mashbean.

**Principal component analysis:** Hotalling[4] developed principal component analysis technique and is concerned with explaining the variance-covariance structure of the set of variables through a few linear combinations of original variables and main objective of principal component analysis is to reduce the dimensionality and interpretation of data set. In principal component analysis, we search for linear combinations that have high correlation with the original variables and orthogonal with each other. Geometrically principal components lie along the perpendicular directions to the axis of the ellipse with maximum variation. Eigenvectors and eigenvalues of variance-covariance matrix determine the direction of maximum variability and variances, respectively.

In practice, we usually concerned with the problem of summarizing the variation in the sample data with n measurements on p variables. Let the random vector:

$$X' = [X_1 \ X_2 \ X_3 \ldots X_p]$$

Follows multivariate distribution with mean vector $\mu$ and covariance matrix $\Sigma$ having pairs of eigenvalues and eigenvectors given as:

$$(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), (\hat{\lambda}_3, \hat{e}_3), \ldots, (\hat{\lambda}_p, \hat{e}_p)$$

Then the ith sample principal component can be determined as:

$$\hat{y}_i = \hat{e}_i X' = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \hat{e}_{i3} x_3 + \ldots + \hat{e}_{ip} x_p \ \forall \ i = 1,2,3,\ldots,p \quad (1)$$

Where:

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \hat{\lambda}_p \geq 0$$

and sample variance of ith sample principal component would be:

$$\text{Var}(\hat{y}_i) = \hat{\lambda}_i, \ \forall \ i = 1,2,3,\ldots,p$$

Also

$$\text{Cov}(\hat{y}_i, \hat{y}_k) = 0, \text{ for } i \neq k$$

A useful visual aid in determining an appropriate number of principal components of data is the scree plot. Scree plot gives us an idea about the contribution of various principal components. An elbow in the scree plot plays an important role in determining the appropriate number of principal components.

**Correlation between original variables and principal components:** Correlation of the variables with the principal components (PC's) often helps to interpret the components. The correlation between ith original variable and jth principal component $y_j$ is given as:

$$r_{\hat{y}_{i,x_k}} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{S_{kk}}} \quad (2)$$

On the basis of correlation coefficient given in equation (2), original variables are classified into three mutually exclusive sets A, B and C. Original variables having positive correlation with PC's are placed in set A, whereas set B and set C contain all variables having negative correlation and no correlation with PC's, respectively. This subdivision is useful for explaining correlation structure.

**Path coefficient analysis:** If the cause and the effect relationship is well defined, it is possible to represent the whole system of variables in the form of a path diagram and path coefficients. Considering mash grain yield (Y) as effect of various causal factors such as plant height ($X_1$), days to flowering ($X_2$), days to first pod maturity ($X_3$), days to 90% maturity ($X_4$), branches per plant ($X_5$), pods per plant ($X_6$), pods length ($X_7$), seeds per plant ($X_8$), 100-seeds weight ($X_9$) and biological yield per plant ($X_{10}$). For explanation purpose, the simple path diagram of effect (Y) and three casual factors say $X_1$, $X_2$ and $X_3$ is shown in Fig. 1.

From the path diagram it is obvious that yield (effect Y) is the result of all X's (casual factors) and other undefined factors designated by "R". Further all the casual factors in turn are correlated. In the Fig. 1 a, b and c are the path coefficients due to respective variables. Path coefficients can be defined as the ratio of standard deviation of the effect due to a given cause to the total standard deviation of the effect. If Y is the effect and $X_i$ is

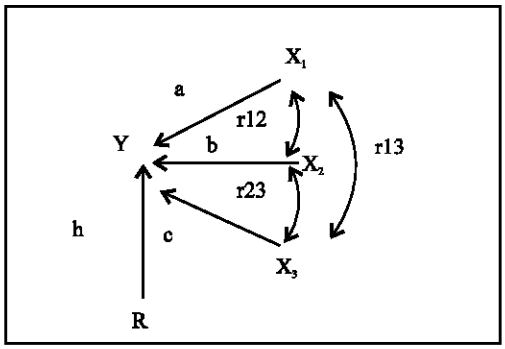


Fig. 1: The simple path diagram

the cause, the path coefficient for the path from cause $X_i$ to the effect Y is given as:

$$\text{Path Coefficient}(X_i \rightarrow Y) = \frac{\sigma_{X_i}}{\sigma_Y} \quad i = 1,2,3,\ldots,10$$

The advantage of the path diagram is that a set of simultaneous equation can be written directly from the diagram and the solution of these equations provides information on the direct and indirect contribution of these casual factors to the effect. Consider the correlation between $X_i$ and Y i.e. $r_{(Xi, Y)}$:

$$r_{(X_i,Y)} = \frac{\text{Cov}(X_i, Y)}{\sqrt{\text{Var}(X_i) * \text{Var}(Y)}} \quad \forall \; i = 1, 2, 3, \ldots, 10 \tag{3}$$

Assuming that:

$$Y = X_1 + X_2 + X_3 + \ldots + X_{10} + R \tag{4}$$

We know that:

$$\bar{Y} = \bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \ldots + \bar{X}_{10} + R \tag{5}$$

By substituting equation (4) and equation (5) in equation (3), we get the equation.

Where, coefficients of $r_{(Xi, Xj)}$ for $j \neq i = 1,2,3 \ldots 10$ are the path coefficients:

$$r_{(X_i,Y)} = \frac{\sigma_{X_1}}{\sigma_Y} + r_{(X_i,X_2)} * \frac{\sigma_{X_2}}{\sigma_Y} + r_{(X_i,X_3)} * \frac{\sigma_{X_3}}{\sigma_Y} + , \ldots + r_{(X_i,X_{10})} * \frac{\sigma_{X_{10}}}{\sigma_Y} \quad \forall \; i = 1, 2, 3, \ldots, 10 \tag{6}$$

Finally we can get a set of simultaneous equations like equation (6) and convert it in matrix form as:

$$A = B * C$$

or

$$C = B^{-1} * A \tag{7}$$

- Where, A is the column vector of correlation between effect i.e. mash grain yield (Y) and each cause i.e. fixed mash plant traits (X's).
- B is the square matrix of pair wise correlation between causes (fixed mash plant traits).
- C is the column vector of Path coefficients.

Steps in path analysis are partitioning the correlation between cause and effect into two parts, which are given as:

(i) Direct effect of $X_i$ on Y, which is $\sigma_{Xi} / \sigma_Y$
(ii) Indirect effect of $X_i$ on Y via remaining $X_j$'s ($i \neq j$) which are given in equation (6).

After having calculated path coefficients, we can obtain the path value for the residual effect (RE), which is estimated as:

$$RE = \sqrt{1 - A^{-1} * C} \tag{8}$$

## RESULTS AND DISCUSSION

Principal component analysis was performed on pooled (with respect to replication) mash data of 37 cultivars, by considering genotypes as cases and 11 traits as variables. By examining correlation matrix between pairs of mash plant traits it was observed that there is a strong correlation between many pair of traits. The strong correlation between traits varies from 0.79 to 0.97. This strong correlation pattern suggests that there is need for reduction in dimensionality.

By applying principal component analysis technique on mash data, the eigenvalues and cumulative variances of correlation matrix of mash plant traits are given in Table 1. It is clear from the Table 1 that the first principal component of mash data accounts 38.6% of the total variability present in the data. While the second principal component accounts 29.5% of the total variability. Also note that the first two principal components accounts 68% of total variability. Similarly first four principal components collectively account 88%.

Now the question of how many principal components to retain can be solved with the help of scree plot.

A useful visual aid to determining an appropriate number of principal components of mash data is the scree

Table 1: Eigen-analysis of correlation matrix

| Principal components | Eigen value | Proportion | Cumulative |
|---|---|---|---|
| $PC_1$ | 4.2447 | 0.386 | 0.386 |
| $PC_2$ | 3.2500 | 0.295 | 0.681 |
| $PC_3$ | 1.2559 | 0.114 | 0.796 |
| $PC_4$ | 0.9281 | 0.084 | 0.880 |
| $PC_5$ | 0.5290 | 0.048 | 0.928 |

plot. An elbow occurs in the scree plot Fig. 2 at about i=4. Consequently the variation in the mash data is summarized very well by first four principal components. So the dimensionality reduces from 11mash plant traits to 4 P.C's. So the first four principal component which are orthogonal with each other and extract maximum of the total variability (about 88%) present in the data are given as:

$$PC_1 = 0.146\,X_1 + 0.255\,X_2 + 0.244\,X_3 + 0.279\,X_4 - 0.372\,X_5 - 0.477\,X_6 - 0.262\,X_7 - 0.119\,X_8 + 0.135\,X_9 - 0.371\,X_{10} - 0.446\,Y$$

$$PC_2 = -0.27\,X_1 - 0.40\,X_2 - 0.44X_3 - 0.43X_4 - 0.15X_5 - 0.12\,X_6 - 0.40X_7 - 0.24X_8 - 0.18X_9 - 0.29X_{10} - 0.15Y$$

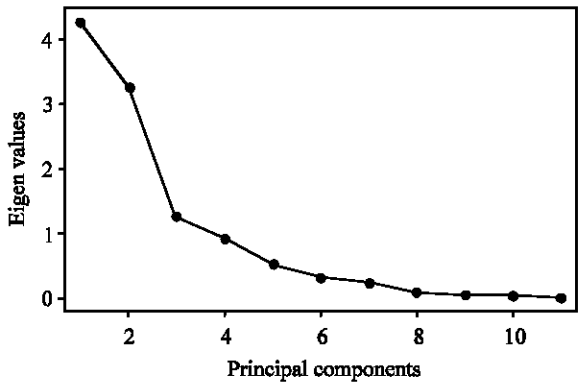


Fig. 2: Scree plot of principal components

$$PC_3 = 56\,X_1 - 0.32\,X_2 - 0.25X_3 - 0.18X_4 - 0.22X_5 - 0.08X_6 + 0.07X_7 + 0.11X_8 + 0.64X_9 + 0.14X_{10} + 0.001\,Y$$

$$PC_4 = -0.05\,X_1 - 0.16\,X_2 - 0.04X_3 - 0.09X_4 - 0.33X_5 - 0.22X_6 + 0.51X_7 + 0.62X_8 - 0.32X_9 - 0.26X_{10} - 0.07Y$$

On the basis of correlation between original variables and principal components given in Table 2, the original variables are divided into three sets A, B and C for positive, negative and no correlation, respectively. Testing of significance procedure at $\alpha$= 5% was made for each correlation between original variables and principal components so as to test for positive, negative or zero correlation and classified in Table 3.

In Table 3, for the first principal component there exists a positive relationship within set A and within set B, while the correlation between each possible pair of members of set A and set B is non significant i.e. uncorrelated. So first principal component represents a contrast between set A and set B. In first principal component, plant height ($X_1$), days to flowering ($X_2$) days to first pod maturity ($X_3$), days to 90% maturity ($X_4$) and 100 seeds weight ($X_9$) have positive correlation between themselves i.e. varies in the same direction. Also in set B traits such as seeds/pod ($X_8$), branches/plant ($X_5$), pod/plant ($X_6$) and biological yield/plant ($X_{10}$) have negative relation with first principal component and positive relation between themselves. For the second principal component, set A is empty. So second principal component is the weighted average of plant height ($X_1$), days to flowering ($X_2$), days to first pod maturity ($X_3$), days to 90 % maturity ($X_4$).

Branches/plant ($X_5$), pod length ($X_7$) seeds/pod ($X_8$), 100-seed weight ($X_9$) and biological yield/plant ($X_{10}$). Now for third principal component there is significant positive correlation within members of set A and members of set B but traits between set A and set B are non-significant i.e., uncorrelated. So the third principal component also represents a contrast between set A and set B. Thus as plant height ($X_1$) increases, 100 seed weight ($X_9$) also increases and as days to flowering ($X_2$) increases, an increase will occur in days to first pod maturity ($X_3$). For the forth principal component the traits within set A is significantly positively correlated but members within set B are uncorrelated. There is highly significant positive correlation between pods/plant ($X_6$) and branches/plant ($X_5$), i.e. as branches/plant increases, pods/plant also increases. Also as branches/plant ($X_5$) increases, seeds/pod ($X_8$) also increases due to positive correlation. But remaining traits of set A and set B is pair wise uncorrelated with each other's. All the remaining principal components are meaningless because they contribute very little in the total variability.

Table 2: Correlation between original traits and principal components

| Traits | $PC_1$ | $PC_2$ | $PC_3$ | $PC_4$ | $PC_5$ | $PC_6$ | $PC_7$ | $PC_8$ | $PC_9$ | $PC_{10}$ | $PC_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 0.30 | -0.48 | 0.62 | -0.05 | -0.52 | -0.03 | -0.10 | 0.05 | 0.02 | 0.02 | 0.00 |
| $X_2$ | 0.53 | -0.72 | -0.35 | -0.15 | -0.01 | 0.05 | 0.12 | 0.17 | -0.11 | -0.02 | 0.01 |
| $X_3$ | 0.50 | -0.80 | -0.28 | -0.04 | 0.02 | 0.01 | -0.02 | -0.08 | 0.09 | 0.08 | 0.07 |
| $X_4$ | 0.57 | -0.77 | -0.20 | -0.08 | 0.00 | 0.03 | 0.05 | -0.09 | 0.06 | -0.05 | -0.09 |
| $X_5$ | -0.77 | -0.28 | -0.25 | -0.32 | 0.04 | 0.17 | -0.37 | 0.04 | 0.00 | 0.01 | -0.02 |
| $X_6$ | -0.92 | -0.22 | -0.08 | -0.21 | -0.09 | -0.08 | 0.11 | 0.05 | 0.11 | -0.11 | 0.03 |
| $X_7$ | -0.25 | -0.72 | 0.08 | 0.49 | 0.17 | -0.34 | -0.17 | 0.00 | -0.04 | -0.03 | 0.00 |
| $X_8$ | -0.54 | -0.43 | 0.12 | 0.59 | 0.01 | 0.39 | 0.08 | 0.02 | 0.02 | 0.00 | 0.00 |
| $X_9$ | 0.28 | -0.32 | 0.71 | -0.30 | 0.46 | 0.08 | 0.02 | 0.03 | 0.02 | -0.01 | 0.01 |
| $X_{10}$ | -0.76 | -0.51 | 0.16 | -0.25 | -0.08 | 0.03 | 0.10 | -0.16 | -0.13 | 0.00 | 0.02 |
| Y | -0.92 | -0.27 | 0.00 | -0.07 | 0.04 | -0.16 | 0.17 | 0.07 | 0.04 | 0.12 | -0.04 |

Table 3: Subdivision of the original variables on the basis of correlation structure

| P.C # | Set A | Set B | Set C |
|---|---|---|---|
| 1 | $X_1, X_2, X_3, X_4, X_9$ | $X_8, X_6, X_{10}, Y$ | $X_7$ |
| 2 | $\phi$ | $X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_9, X_{10}$ | $X_6, Y$ |
| 3 | $X_1, X_9$ | $X_2, X_3$ | $X_4, X_5, X_6, X_7, X_8, X_{10}, Y$ |
| 4 | $X_6, X_8$ | $X_5, X_9$ | $X_1, X_2 X_3, X_4, X_7, X_{10}, Y$ |

**Path analysis:** was performed on mash plant data with mash grain yield (Y) as response variable. In this analysis, Y stands for grain yield, while 1, 2, 3, 4, . . . 10 stand for plant height, days to flowering, days to first pod maturity, days to 90% maturity, branches per plant, pods per plant, pods length, seeds per plant, 100-seeds weight and biological yield per plant, respectively. The "B" matrix is for the pair wise correlation coefficients between each fixed traits, whereas column vector A is for the pair wise correlation between grain yield (Y) and each of the fixed mash plant traits. Vector C is calculated to determine path coefficients by using the relationship (7).

$$C = \begin{bmatrix} P_{1Y} \\ P_{2Y} \\ P_{3Y} \\ P_{4Y} \\ P_{5Y} \\ P_{6Y} \\ P_{7Y} \\ P_{8Y} \\ P_{9Y} \\ P_{10Y} \end{bmatrix} = \begin{bmatrix} -0.122004 \\ 0.152293 \\ 0.172730 \\ -0.458422 \\ -0.153435 \\ 0.683157 \\ 0.208041 \\ 0.006186 \\ 0.056645 \\ 0.293296 \end{bmatrix}$$

The elements of vector C represent the direct effects of each fixed trait on mash grain yield (Y). Path analysis suggested that pods/plant ($X_6$) has maximum positive direct effect (i.e. 0.6832) on mash grain yield as compared to remaining mash plant traits. It indicates that with the increase in pods/plant ($X_6$), mash grain yield also increases, whereas days to 90% maturity ($X_4$) has maximum negative direct effect i.e. -0.4584 on mash grain yield, which shows in increase in days to 90% maturity, mash grain yield will decrease. Branches /plant ($X_5$) and biological yield/plant ($X_{10}$) have maximum positive indirect effect i.e., 0.54 and 0.58, respectively on mash grain yield via pods/plant. But both days to flowering ($X_2$) and days to first pod maturity ($X_3$) have approximately similar negative indirect effect i.e., -0.43 on mash grain yield via days to 90% maturity.

Finally it is recommended that with the increase in the number of pods per plant, the mash grain yield increases, so the varieties having more pod per plant must be recommended. Also the breeders should be alert about the increase in days to 90% maturity ($X_4$), which cause the reduction of the total mash yield.

**Residual effect:** was calculated from equation (8), indicates how best casual factors accounted the variability in the mash grain yield. It was observed that all the casual factors explained 76% of the variation in the mash grain yield. The reason of low residual effect is that some of the mash plant traits had non-significant relation with mash grain yield.

## CONCLUSION

The objective of the study was to find the regressors's direct–indirect effects on mash grain yield and to find correlation structure through path analysis and principal component analysis, respectively. In principal component analysis it was observed that dimensionality of mash data reduced from 11 traits to only 4 principal components, which contributes 88% of the total variability present in mash data. Also the second principal component was observed as weighted average of plant height ($X_1$), days to flowering ($X_2$), days to first pod maturity ($X_3$), days to 90 % maturity ($X_4$), seeds/pod ($X_8$), pod length ($X_7$), pods/plant ($X_6$), branches/plant ($X_5$), 100-seed weight ($X_9$), biological yield/plant ($X_{10}$) and grain yield (Y).

In path analysis pods/plant showed that with the increase in pods /plant, Mash grain yield also increases, whereas increase in days to 90% maturity, Mash grain yield will decrease. Also branches /plant and biological yield/plant have maximum positive indirect effect i.e., 0.54 and 0.58, respectively on mash grain yield via pods/plant. On the other hand both days to flowering and days to first pod maturity have approximately similar negative indirect effect i.e. -0.43 on mash grain yield via days to 90% maturity.

## REFERENCES


1. Shukla, G.K., 1972. Some statistical aspects of partitioning genotype environment components of variability heredity. Cereal Res., Communications, 29: 237-245.
2. Ayub, K., R. Muhammad, K. Amjad, M.I. Khan and R. Shahid, 2000. Correlation and path coefficient analysis for the yield contributing parameters in *Brassica napus*. Pak. J. Agril. Res., 16: 127-130.
3. Kubsad-VS, S.A. Desai, C.P. Mallapur and G.G. Gulaganji, 2001. Path coefficient analysis in safflower. J. Maharashtra Agril. Univ., 25: 321-322.
4. Hotalling, H., 1936. Relations between two sets of varieties. Biometerika, 28: 321-377.