Studies on the Multivariate Statistics Oriented Genetic Parameters and Their Uses

Muneharu SATO, Anton J. BOCKHOLT* and Teiji KOKUBU
(Laboratory of Plant Breeding)

Received for Publication September 10, 1984

Introduction

Although factor analysis methods have not been widely used, they have a good potential to solve various agronomic and biological problems. However, to utilize the methods in breeding programs, they should be incorporated with genetics. Therefore, the objectives of this study are to evaluate a factor analysis as a breeding tool and to propose new genetic parameters based on a factor analysis.

1. Review of literature

The concept of factor analysis method was founded by mathematicians and psychometrists such as Pearson[12], Spearman[17] and Garnett[8] early in this century. And theoretically, a factor analysis is one of the most powerful methods to extract the independent components from many observed variables and clarify the interrelationships between and among characters.

In the plant breeding related fields, it has been used primarily to analyze yield and quality components. Mehra *et al.*[10] for example, used it to examine the yield components in oat and found that two interactive hormonal systems were controlling the components. Walton[18] applied the technique to determine the relationships among 14 characters of wheat and classified them into four groups. It was also used to isolate the several independent components from 13 bread-making quality parameters and to increase the efficiency of the breeding program[1]. The increase of main branch length, number of primary baranches, stem girth, and number of leaves were suggested to increase fodder yield by the results of a factor analysis on cowpea (*Vigna unguiculata* (Linn) Walp., *V. sinensis* (Linn) Savi.)[16]. Denis and Adams[5] reported that weight, number of leaves and plant archtecture were consistently isolated across and within locations and in various subsets of the data by applications of a principal factor analysis on 22 morphological and yield-determining characters of 16 cultivars and strains of dry beans (*Phaseolus vulgaris* L). A factor analysis on 15 agronomically important characters of maize (*Zea mays* L.) revealed that grain yield and qualitative characters were controlled by separate factor components[11].

2. Proposed method

(1) The basic concept

Generally speaking, breeders have to depend on phenotypically expressed characters to evaluate a plant. However, little is known about the relationship between genes and quantitative characters. Characters are usually chosen rather artificially for our conveniences of observations first, then the

* Department of Soil and Crop Sciences, Texas A & M Univ., College Station Texas, U. S. A.

corresponding genes are deduced. One-to-one correspondence between a gene and a particular character, however, might not always exist. The presence of highly correlated characters indicates that some characters are controlled by a similar set of gene expression systems and physiological regulation systems. It is, therefore, conceivable that what we call a "character" represents only one facet of such systems. In fact, it is known that pleiotropic genetic systems and linkage systems are operating in plants. And the complex gene regulation systems of Eucaryotes were also presented by Davidson and Britten[4].

Factor analysis methods have been developed to reduce the dimension of original observations and extract the common, usually independent, components. And further, recently, it was also suggested that a factor analysis can handle the multicollinearity problems and the difficulties caused by erroneous measurements, both of which are rather common in the plant breeding field[13,14,15].

In the pleiotropic genetic system, a single gene would be expressed in two or more characters. The linkage system also correlates two different characters. Some clusters of genes might be expressed as several phenotypically observable characters through interrelated complex physiological regulation systems. Mathematically, the relationship between genes and expressed characters can be interpreted as a mapping. If the simple one-to-one mapping system is detected between the common factor and the observed characters, the factor analysis would reduce the two dimensional observed data to the one dimensional common factor. Considering the gene expression systems known at present, the chance is that the isolated factor might be the genetic-physiological system common to the two characters originally observed. Then what the factor analysis has done here is to establish the one-to-one correspondence between the genetic-physiological systems and the extracted common factor.

Those facts stated above suggest that a factor analysis could be an effective tool to find out the interrelationships between characters, as well as the possible independent gene expression systems from many observed characters.

(2) Factor analysis oriented genetic parameters

Assuming a factor analysis is capable of isolating the independent gene expression systems, it may be incorporated with genetics by defining genetic parameters for the extracted genetic components.

The model of a factor analysis is:

$$X = YA' + U \quad \cdots(1),$$

where X is a p×n data matrix (usually in standard unit); Y is a m×n factor score matrix; A is a p×m factor loading matrix; U is a p×n uniqueness matrix; and $Y'U=O$, $Y'Y=I$, $U_i' U_j = O$ $(i \neq j)$, where U_i and U_j are the i-th and the j-th column vector of the matrix U, respectively, and the symbol $'$ denotes the transposed matrix.

Here, Y is an orthogonally transformed matrix with reduced dimension and is considered to represent the extracted common factors. The i-th column of Y indicates i-th factor.

By exercising the same analogue to estimate a conventional heritability of a character, it would be possible to obtain the additive genetic poition of the total variance of each factor, or the "factor score heritability" of each factor. For example, a factor score is analysed by the following model:

$$Y_{ijk} = \mu + m_i + f_{j(i)} + e_{ijk},$$

where Y_{ijk}=the factor score of the k-th observation of the offspring from the j-th female parent crossed with the i-th male; μ=a common effect for the whole population; m_i=the effect of the i-th male parent; $f_{j(i)}$=the effect of the j-th female parent within the i-th male parent; and e_{ijk}=the random error.

Then, the "sib correlation of factor score" is estimated as $\hat{t}$ (factor) $= V_m/V_t$ or the estimated factor score heritability as $\hat{h}^2_{FA}=4V_m/V_t$, where V_m represents the estimate of variance component of male parent and V_t shows the total variance component of the factor. If the original data consist of parent-offspring measurements, the regression method can be used to obtain the factor score heritability by applying the same logic to the factor score as used in finding the heritability of a character. And if the half-sib population is used for the analysis, by multiplying the sib-correlation by four the factor score oriented heritability will be estimated.

Essentially, factor score oriented heritabilities are the heritabilities of the common components extracted by a factor analysis, representing the additive genetic portion of the independent gene expression systems. Therefore, if the gene expression system for a particular character is independent of others, the estimate of factor score oriented heritability and that of conventional heritability should be alike. If the gene expression system is complex, the factor score oriented heritability would make more sense than the conventional one which may lack the one-to-one corresponding set of genes.

To make the differences between conventional genetic parameters and the factor anlysis oriented ones clear, Fig. 1 is presented. As shown here, if the character of the parent (X_p) has the gene or genes to express it and transmit them to the offspring to express the character (X_0) directly as in case 1, both estimates would provide the same acceptable values. But if X is controlled by the results of the expressions of Y or a group of many other characters as in case 2 and case 3 and has no direct path to the corresponding character of the offspring (X_0), the conventional heritability estimates may become somewhat ambiguous and prone to errors. However, in the factor analysis oriented method, first those belonging to the same system are clustered into one group, then the amount of the additive genetic variance of the group is computed. Hence, the reasonable value could be obtained in the factor analysis oriented method.

(3) A factor analysis oriented selection index

As the extracted components are mutually independent, the selection index, which indicates the relative aggregated genetic performance, can be derived as follows.

Let the diagonal elements of $\boldsymbol{B}$ matrix with the dimension of m×m be the factor score heritabilities and let its off-diagonal elements be zero. Then the factor scores of the offspring can be estimated from those of parents by:

$$Y_0 = Y_c + Y_p B,$$

where Y_0=the p×m estimated factor score matrix of offspring; Y_c=the p×m constant matrix, the column elements of which are identical within the column; and Y_p=the p×m factor score matrix of parent.

Then, from the model of a factor analysis (formula (1)),

$$X_0 = Y_0A' + U = Y_cA' + Y_pBA' + U = Y_pBA' + (Y_cA' + U),$$

where X_0 is the p×m estimated data matrix of offspring.

Now, Y_cA' are constant to all estimated offspring data and U matrix is unknown and can be considered to be an error-like term. Hence, for relative comparison of performance of the individual to others, only the Y_pBA' term is important. Therefore, the estimates of aggregated genotypic performance of the individuals to be selected in the population, or the index vector will be:

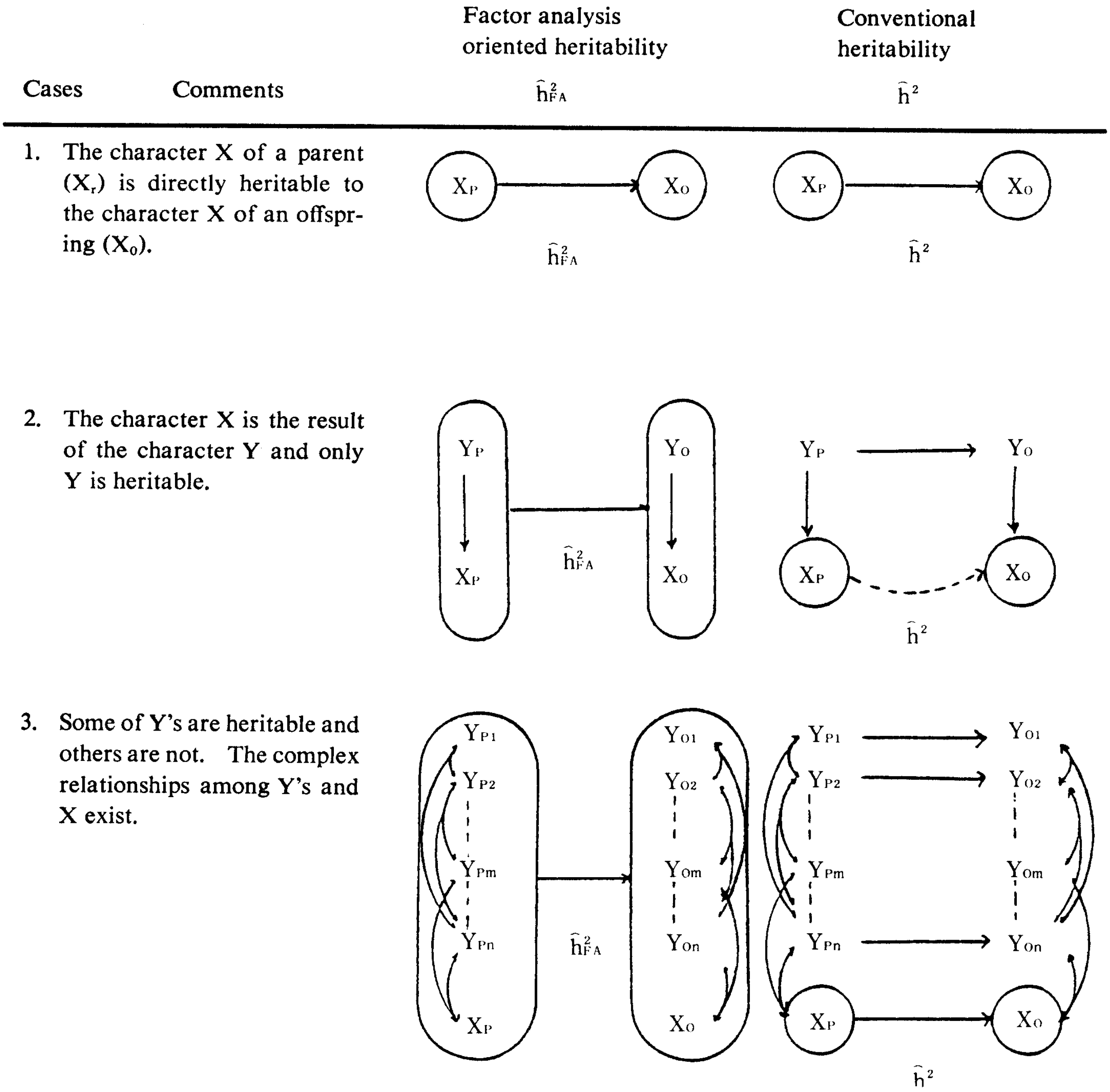
Factor analysis oriented heritability
Conventional heritability
Cases
Comments
$\hat{h}^2_{FA}$
$\hat{h}^2$
1. The character X of a parent (X_r) is directly heritable to the character X of an offspring (X_0).
X_P
X_O
$\hat{h}^2_{FA}$
$\hat{h}^2$
2. The character X is the result of the character Y and only Y is heritable.
Y_P
Y_O
X_P
X_O
$\hat{h}^2_{FA}$
$\hat{h}^2$
3. Some of Y's are heritable and others are not. The complex relationships among Y's and X exist.
Y_{P1}
Y_{P2}
Y_{Pm}
Y_{Pn}
X_P
Y_{O1}
Y_{O2}
Y_{Om}
Y_{On}
X_O
$\hat{h}^2_{FA}$
$\hat{h}^2$

Fig. 1. Diagramatic representation of heritability estimates derived from factor analysis and conventional univariate analyses.

$$\text{Index}=Y_p\boldsymbol{B}\boldsymbol{A}'\boldsymbol{w},$$

where Y_p denotes the row vector of factor score of the individual; $\boldsymbol{B}$ is the diagonal matrix, and the diagonal elements of which are the factor score heritabilities; $\boldsymbol{A}$ represents the factor loading matrix; and $\boldsymbol{w}$ is the economic vector for the standardized phenotypic character measurements.

As the factor score is independent of each other and the off-diagonal elements of the matrix $\boldsymbol{B}$ are zero, the constant multiplication factors such as "four" for the heritability estimates from the half-sib correlations, have nothing to do with the relative value of the index. Therefore, the ranking ordered by the indices obtained by the factor score intraclass correlations and by the factor score heritabilities should be identical. Hence, the diagonal elements of the matrix $\boldsymbol{B}$ can be the factor score intraclass correlation vector in practice.

Materials and Methods

To evaluate the validity of the factor analysis and the proposed method, the data of yield related characters of sorghum (*Sorghum bicolor* (L.) Moench) originally collected for the Texas Agricultural Experiment Station project TAES H-1904 by Mr. G. Perez and Dr. F. R. Miller were analysed. The data consist of two sets of subpopulations: the male lines of one subpopulation (G1) are RTx 430 and 76CS256, while those of the other (G2) are 76CS4409 and RTAM28. The female parental lines of both subpopulations are A1388, ATx3197, ATx7801, A1388 × ATx3197, A1388 × BTx7801 and ATx3197 × BTx7801. The sample sizes of G1 and G2 are 889 and 847, respectively. The measurements on height (HT), days to 50 percent flowering (DAY), head length (HED), head exsertion (EX), midge damage score (DMID), and grain yield (YLD) were used in this study.

The summary statistics of the measurements of G1, G2 and their pooled population (Pooled) are shown in Table 1. Although 50 percent of the genes of G1 and G2 are the same, these two subpopulations are significantly different at the one percent level by all Hottelling's-Lawley trace, Pillai's trace, Wilks' criterion and Roy's maximum root criterion multivariate analysis of variance

Table 1. Summary statistics of the data used for this study

	Populations								
	G1			G2			Pooled		
Sample size	889			847			1736		
Characters	Mean $\bar{x}$	Variance s^2	Coefficient of variation c.v. (%)	Mean $\bar{x}$	Variance s^2	Coefficient of variation c.v. (%)	Mean $\bar{x}$	Variance s^2	Coefficient of variation c.v. (%)
DAY	74.65	19.30	5.88	74.73	19.77	5.95	74.69	20.25	6.02
HT (cm)	128.84	110.24	8.15	126.29	159.07	9.88	127.60	135.96	9.14
HED (cm)	30.78	14.83	12.51	29.62	14.73	12.96	30.21	17.56	13.87
EX (cm)	13.54	43.98	48.98	16.35	45.39	41.21	14.91	54.61	49.56
DMID	11.07	116.65	97.52	10.55	97.99	93.80	10.82	114.70	98.98
YLD (g/plant)	43.82	257.16	36.59	37.28	153.57	33.24	40.63	219.34	36.45

Phenotypic correlation coefficients

	DAY	HT	HED	EX	DMID	YLD	
DAY	1.	0.20**	0.15**	0.20**	0.05	0.03	— G1
		0.16**	0.27**	0.25**	0.14	−0.17*	— G2
HT		1.	−0.03	0.50**	−0.13**	0.05	— G1
			0.12**	0.40**	−0.02	0.07*	— G2
HED			1.	−0.42**	0.34**	0.28**	— G1
				−0.41**	0.30**	0.40**	— G2
EX				1.	−0.21**	−0.39**	— G1
					−0.22**	−0.31**	— G2
DMID					1.	−0.11**	— G1
						−0.03	— G2
YLD						1.	

*, ** Significant at the 0.05 and 0.01 levels, respectively.

tests and also Hotelling's T test. Univariate t-tests also showed that differences between population means were highly significant in all characters but in DAY and DMID.

Populations, G1, G2 and Pooled, were analysed by an iterated principal axis factor analysis with the variamax rotation, independently. And a maximum likelihood factor analysis was also attempted for Pooled population to study the difference of the analysis methods.

The factor scores obtained by the principal factor analysis were used to find the proposed genetic parameters or factor score oriented sib correlations. At the same time, the conventional heritabilities of all six characters were calculated for these three populations to make a comparison between the proposed and the conventional methods in their effectiveness.

Results and Discussion

The results of the factor analysis on three populations of sorghum are presented in Tables 2 and 3. As in all analyses, the first three components extracted account for 100 percent of the variance of common factors of six originally observed characters before rotations, the first three factors were retained for further analyses. Despite the differences of the populations and analysis methods, both factor patterns and communality estimates resemble each other. As the results are those of rotated factor analyses, the factor contributions are well balanced. Although the communality estmates of

Table 2. Rotated factor patterns of five sorghum populations

Factors	Variables	Interated principal axis factor analysis			Maximum likelihood factor analysis
		G1	G2	Pooled	Pooled
1	DAY	0.04	−0.14	−0.05	−0.03
	HT	0.03	0.06	0.05	0.05
	HED	0.29	0.41	0.35	0.41
	EX	−0.46	−0.32	−0.38	−0.47
	DMID	−0.14	0.02	−0.04	−0.03
	YLD	0.83	1.00	1.00	0.99
	Proportion*	(31%)	(36%)	(36%)	(38%)
2	DAY	0.12	0.11	0.11	0.15
	HT	1.00	0.96	1.00	1.00
	HED	−0.06	0.06	0.01	0.01
	EX	0.51	0.48	0.46	0.47
	DMID	−0.14	−0.04	−0.09	−0.09
	YLD	0.02	0.01	0.02	0.04
	Proportion	(40%)	(33%)	(35%)	(35%)
3	DAY	0.28	0.55	0.40	0.52
	HT	0.05	0.08	0.07	0.06
	HED	0.63	0.58	0.61	0.48
	EX	−0.45	−0.58	−0.54	−0.53
	DMID	0.51	0.37	0.42	0.40
	YLD	0.02	−0.04	−0.03	−0.12
	Proportion	(29%)	(31%)	(29%)	(27%)

* The percentage of the factor contribution.

Table 3. The communality estimates for six variables in sorghum

Variables	Iterated principal axis factor analysis			Maximum likelihood factor analysis
	Gl	G2	Poolded	Pooled
DAY	0.10	0 34	0.34	0.29
HT	1.00	0.93	1.00	1.00
HED	0.49	0.50	0.49	0.40
EX	0.68	0.68	0.64	0.72
DMID	0.30	0.14	0.19	0.17
YLD	0.70	0.99	1.00	1.00

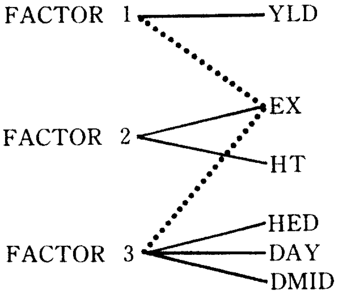

Fig. 2. Arrow diagram showing the effects of factors on six variables in sorghum. The solid lines designate strong positive relationships, while the broken lines indicate negative relationships.

midge damage score (DMID) of all populations are very low, plant height (HT) and grain yield (YLD) are fully explained by the extracted factor components.

Among six characters studied, YLD seems to be relatively independent of others and accounts for the most of the first isolated component, Factor 1 (Fig. 2). On the other hand, head length (HED), floweing day (DAY) and DMID displayed their complex relationships as represented by Factor 3. Head exertion (EX) is influencing all extracted factors in different ways. For instance, its relationship with Factor 2 is positive while its associations with others are negative.

As yield is relatively independent from other characters, it might be effective to select for yield alone. However, it should be better to use a selection index since the negative influence from exertion and the indirect effects of other characters on yield are suspected.

The male half-sib correlation estimates of three factor scores and six characters originally observed are compared in Tables 4 and 5. Although different scoring coefficients were applied to estimate factor scores, the relatively consistent results of the proposed method were obtained, suggesting an interchangeability of coefficients. Factor 2 represented by HT and part of EX is not highly heritable, while Factor 3 having a rather complex structure as presented in Fig. 2 showed a relatively high proportion of additive components. The similarity of the results among populations was observed in all extracted factors but Factor 1, which is mainly contributed by yield.

The results obtained by the conventional method, however, were not consistent. If an insignificant amount of nonadditive genetic effects and inbreeding effects are assumed, the conventional heritability estimates will be obtained by multiplying the sib correlations by four. As compared with the results of experiments by Eckbil *et al.*[6] and Crook and Casady[3], the estimates of DAY, HT, and YLD obtained in this study are smaller.

Table 4. Male half-sib correlation estimates of common factors extracted from six characters of sorghum

Populations	Methods* I			II			III		
	Factors			Factors			Factors		
	1	2	3	1	2	3	1	2	3
G1	2.81	1.20	22.93	1.86	1.25	21.15	2.18	0.88	19.36
G2	2.54	0.0	20.00	0.0	0.0	24.83	0.93	0.01	14.65
Pooled	6.29	0.58	18.19						

* The estimates were computed from the factor scores, the scoring coefficients of which were those of pooled population, of their own, and of the other population in Methods I, II, and III, respectively.

Table 5. Male half-sib correlation estimates of six characters of sorghum

Populations	Characters DAY	HT	HED	EX	DMID	YLD
G1	0.	1.30	35.77	0.	18.44	2.70
G2	2.33	0.	0.	39.45	0.	0.
Pooled	0.	0.94	15.50	20.90	5.71	6.86

The estimate of Factor 1 and that of YLD within a population are similar, since YLD is essntially independent of other characters and Factor 1 is mainly represented by YLD as found by factor analyses. On the other hand, the conventional heritability estimates of other complexly interrelated characters appear to be rather inconsistent. The results show the fact that the more the gene expression systems become complex, the more the conventional heritability estimates come to be prone to error. The factor score oriented heritabilities, on the other hand, seems to be less subject to fallacies, because they are supposed to represent the additive varaiance portion of independent gene expression systems extracted by a factor analysis. At least the proposed geneitc parameters would give the additional information on genetic structures of a crop to breeders.

Judging from the results obtained, the proposed parameters appeared to be as good as, if not better than, the conventional heritabilities. Further, it can confer a new insight to the genetic analysis of a plant by providing the genetic informations of independent components, which possibly represent gene expression systems.

The proposed method could be applied to a wide range of fields. In a selection program, for instance, it would help to pick up one or more characters contributing to the genetic progress most, even if the genetic analyses of the isolated factors may not be the primal interest. The very detection of independent biological systems ought to be beneficial for the breeders. Further, a factor analysis should reduce the problems associated with multicollinearity as well as erroneous measurements when some regression analyses are desired.

Multivariate analysis methods have been proven to be effective to evaluate the genotype × environment interactions[2,7,9]. By combining those multivariate analysis methods and the proposed method, it might be possible to ascertain what kind of system responds to the change of the environment. Then, the mechanism of the genotype×environment interaction could be assessed with less difficulty than by the conventional univariate analysis oriented methods.

The contributions to basic biological studies such as plant physiology, genetics, plant pathology, etc. are also expected. For example, the comparisons made among the isolated factors, their genetic parameters, and the degree of the resistance of a plant to a disease or an insect would reveal the resistant mechanism. At least, it might show a relatively easy way to identify the resistant mechanism. In general, the multivariate analysis oriented method could provide the information on the genetical and physiological dependency of the characters more clearly than the univariate techniques. Therefore, both genetic and physiological studies could be made simpler by the application of multivariate analysis methods and the proposed method than the conventional methods.

Although the results seemed to be favorable to the proposed method, both the number of characters analyzed and the sample sizes of populations used in this study were too small to draw definit conclusions. Therefore, more detailed studies and carefully controlled experiments should be executed before the confirmation the validity of the proposed technique.

Summary

Multivariate statistics methods have rarely been used by plant breeders. However, recent advancements of electronic computers have virtually eliminated the problems associated with complicated calculations required to execute a multivariate analysis. And, in fact, the number of applications of multivariate analyses in various disciplines of science has been growing rapidly, recently.

Therefore, in this study, attention is paid primarily to evaluate one of the multivariate statistical methods, Factor Analysis, and to propose a new concept of hereditary parameters and a new selection index.

Two sorghum (*Sorghum bicolor* (L.) Moench) populations and their pooled populations were analysed by two methods of factor analyses and the proposed genetic parameters were derived.

Six original variables, grain yield, head exertion, plant height, head length, flowering day and midge damage score were reduced to three major factors which account for almost all of their variations. Factor patterns for these six variables of sorghum indicated that grain yield might relatively be independent of other characters. The proposed factor score oriented genetic parameters revealed that Factor 2 represented mainly by plant height and the part of head exertion is not a strongly heritable factor.

The results of this study indicated that the proposed method might provide new information to breeders and that it ought to be more reliable than the conventional univariate oriented one. It was also suggested that factor analyses would be acceptable in plant breeding programs. However, as the number of the variables analysed and the population sizes of data used in this study were relatively small for this kind of study, a more detailed evaluation should be done before full scaled field applications.

Acknowledgements

The authors want to express their gratitude to Dr. F. R. Miller and Mr. G. Perez for providing the authors with their data sets.

References

1) Briggs, K. G. and Shebeski, L. H.: An application of factor analysis to some breadmaking quality data. *Crop Sci.*, **12**, 44–46 (1972)

2) Byth, D. E., Eisemann, R. L. and DeLacy I. H.: Two-way pattern analysis of a large data set to evaluate genotypic adaptation. *Heredity*, **317**, 215–230 (1976)

3) Crook, W. J. and Casady, A. J.: Heritability and interrelationships of grain-protein content with other agronomic traits of sorghum. *Crop. Sci.*, **14**, 622–624 (1974)

4) Davidson, E. H. and Britten, R. J.: Organization, transcription and regulation in the animal genome. *Quart. Rev. Biol.*, **48**, 565–613 (1973)

5) Denis, J. C. and M. W. Adams.: A factor analysis of plant variables related to yield in dry beans. I. Morphological traits. *Crop Sci.*, **18**, 74–78 (1978)

6) Eckebil, J. P., Ross, W. M., Gardner, C. O. and Maranville, J. W.: Heritability estimates, genetic correlations and predicted gains from S1 progeny test in three grain sorghum random-mating populations. *Crop Sci.*, **18**, 373–377 (1977)

7) Freeman, G. H. and Dowker, B. D.: The analysis of variation between and within genotypes and environments. *Heredity*, **30**, 97–109 (1973)

8) Garnett, J. C. M.: On certain independent factors in mental measurement. *Proc. Roy. Soc. London*, **96**, 91–111 (1919)

9) Grafius, J. E. and Kiesling, R. L.: The prediction of the relative yields of different oat varieties based on known environmental variables. *Agron. J.*, **52**, 396–399 (1960)

10) Mehra, K. L., Mal, B., Screenath, P. R., Magoon, M. L. and Katiyar, D. S.: Factor analysis of fodder yield components in oats. *Euphytica*, **20**, 590–601 (1971)

11) Motto, M.: Heritability and interrelation of seed quality and agronomic traits in a modified opaque-2 synthetic variety of maize (*Zea mays* L.). *Maydica*, **XXIV**, 193–202 (1979)

12) Pearson, K.: On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, *ser* **2**, **G**, 559–572 (1901)

13) Scott, J. T., Jr.: Factor analysis and regression. *Econometrica*, **34**, 552–562 (1966)

14) Scott, J. T., Jr.: Factor analysis regression revised. *Econometrica*, **37**, 719 (1969)

15) Scott, J. T., Jr. and Fleishman, A.: Statistical analysis of the goodness of classical factor analysis regression (CFAR). *Agric. Exp. Station Bull.*, **759**, Agric. Exp. Station, College of Agric., Univ. of Illinois at Urbana-Champaign (1978)

16) Singh, C. B., Mehra, K. L., Kohli, K. S. and Magoon, M. L.: Correlations and factor analysis of fodder yield components in cowpea. *Acta. Agron. Acad. Sci. Hungaricae*, **26**, 378–388 (1977)

17) Spearman, C.: General intelligence, objectively determined and measured. *Am. J. Psych.*, **15**, 201–293 (1904)

18) Walton, P. D.: Factor analysis of yield in spring wheat (*Triticum aestivum* L.). *Crop Sci.*, **12**, 731–733 (1972)