Arunachalam, V. and Murty, B. R.
Factor Analysis of Diversity in the Genus Sorghum.
Indian Journal of Genetics and Plant Breeding, 27 (1967) : 123-135.

[REPRINTED FROM THE *Indian Journal of Genetics and Plant Breeding*, Vol. 27, No. 1 March, 1967. Pages 123-135]

FACTOR ANALYSIS OF DIVERSITY IN THE GENUS *SORGHUM**

B. R. MURTY and V. ARUNACHALAM

Division of Genetics, Indian Agricultural Research Institute, Delhi-12

(Accepted : 9–ii–1967)

Factor analysis as a branch of multivariate analysis useful to explain the inter-correlations of variables is well known (Holzinger and Harman, 1941 ; Maxwell, 1961; Thurstone, 1947; Lawley and Maxwell, 1963; Rao, 1964). It helps to find out the number and nature of causative influences on which more intensive work can be concentrated. As pointed out by Cattell (1965), its utility lies not only at the exploratory stages of research but also at later stages where the simultaneous action of several factors influencing a variable is to be critically analysed. While the principal component analysis breaks down a covariance matrix into a set of orthogonal components equal in number to the number of variates irrespective of the distribution of the variates or even their randomness, a factor model assumes that the p correlated variables follow a multivariate normal distribution and that their inter-correlations can be adequately accounted for by k factors ($k<p$) which are linear and additive (Maxwell, 1961; Rao, 1964). Thus, in factor analysis, the matrix of covariances can be explained by a smaller number of hypothetical variates or factors. Such an approach is important in studies on biological evolution where the experimenter is unlikely to have *a priori* knowledge of the causal influences.

A multivariate analysis of genetic divergence in the genus *Sorghum* (wild and cultivated forms) using quantitative characters related to fitness under natural and human selection revealed that D^2 statistic and principal component analysis were powerful enough to differentiate not only between species but also between the species and their hybrid derivatives (Chandrasekariah, 1964). It was felt useful, therefore, to obtain the factors responsible for differentiation among species using the same material.

Subsequent to the above study, data became available on 80 elite populations from a world collection of nearly 10,000 genetic stocks. These 80 populations were selected for their high grain productivity and were, therefore, products of intense human selection in cultivated *Sorghums*. This provided an opportunity of utilizing factor analysis to compare the causal influences under *natural* and *human* selection for the diversity found in this genus.

*Presented at the 20th annual general meeting of the Indian Society of Agricultural Statistics held at Waltair in January, 1967.

Material and Methods

The material consisted of (A) 46 populations of Eu-Sorghums representing 22 species (according to the classification by Snowden, 1936) and (B) 80 elite populations from 16 countries forming a representative sample of a world collection. The material A and B were grown during July–December 1963 and 1965 respectively in a randomised complete block design at the Division of Genetics, Indian Agricultural Research Institute, Delhi. Observations on ten characters related to fitness were taken on random samples of five plants during 1963 and on twelve characters on random samples of three plants during 1965. Genotypic and environmental correlation matrices formed for both the years are presented in Tables 1 and 2.

Table 1

Genotypic and environmental correlation matrix of 10 characters in Sorghum, *1963*

	1	2	3	4	5	6	7	8	9	10
				Genotypic						
1	·349									
2	−·400	·231								
3	·183	−·260	·883							
4	−·220	−·036	−·212	·435						
5	·069	−·200	·883	·008	·885					
6	·341	−·155	·345	−·250	·196	·345				
7	−·139	·206	·312	·435	·525	·056	·554			
8	−·002	−·247	·883	−·037	·885	·129	·406	·885		
9	·349	−·458	·208	−·478	−·033	·094	−·529	·013	·349	
10	−·240	·231	·194	·137	·193	·077	·554	·058	−·154	·554
				Environmental						
1	·281									
2	−·426	·335								
3	·212	·276	·758							
4	·140	−·028	·332	·787						
5	·281	·335	·758	·424	·758					
6	·143	−·021	·140	−·022	·179	·228				
7	·068	·091	·466	·298	·584	·091	·584			
8	·143	·007	·490	·787	·530	·181	·454	·787		
9	−·019	·115	·284	−·178	·158	·104	·221	−·127	·284	
10	·273	−·338	·331	·217	·306	·228	·565	·379	·127	·565

1. Growth rate; 2. Days to flower; 3. Panicle length; 4. Number of primaries per node on the panicle; 5. Length of primaries; 6. Angle of primaries; 7. Number of secondaries; 8. Distance between whorls; 9. Distance within whorls; 10. Number of fertile spikelets.

TABLE 2

Genotypic and environmental correlation matrix of 12 characters in Sorghum, *1965*

	1	2	3	4	5	6	7	8	9	10	11	12
					*Genotypic**							
1	—											
2	—	·897										
3	—	·897	·897									
4	—	·717	·773	·773								
5	—	·652	·613	·530	·652							
6	—	−·441	−·489	−·491	−·520	·523						
7	—	·041	·012	−·113	·001	·095	·966					
8	—	·205	·296	·215	·312	−·093	·308	·448				
9	—	·139	·091	·056	·200	−·110	·455	·102	·778			
10	—	−·074	−·121	−·272	−·045	·278	·966	·448	·384	·966		
11	—	−·023	−·018	−·054	·079	−·084	·562	−·031	·778	·494	·778	
12	—	−·250	−·286	−·388	−·303	·523	·460	−·153	·502	·567	·721	·721
					Environmental							
1	·198											
2	−·068	·058										
3	·044	−·108	·327									
4	·107	·007	·327	·327								
5	·034	·011	·201	·158	·287							
6	−·027	·026	−·171	−·003	−·093	·586						
7	−·127	−·010	−·049	−·036	·287	·028	·287					
8	·198	−·011	−·106	−·074	·033	·044	·044	·198				
9	−·148	−·093	·009	−·031	·093	·002	·193	−·012	·596			
10	−·067	·058	−·025	−·087	·087	·074	·172	−·116	·048	·172		
11	−·065	·044	·027	−·035	−·024	·056	·125	·093	·596	·031	·596	
12	·012	−·046	−·018	·161	·001	·586	·050	·034	·073	·036	·146	·586

1. Rate of emergence; 2. Days to 50 per cent. flower; 3. Number of leaves per plant; 4. Stem diameter; 5. Height of plant; 6. Number of productive tillers; 7. Length of panicle; 8. Number of whorls on rachis; 9. Breadth of panicle; 10. Length of rachis; 11. Weight of panicle; 12. Grain yield per plant. *Character 1 was not included due to the high environmental variation.

The highest correlation coefficient in each array was taken as an estimate of the array communality as suggested by Cattell (1965). The factors were obtained by the centroid method which provides adequate solutions (Maxwell, 1961). The variables were reflected in the origin, when necessary, to remove the centroid from the origin in the residual factor space and to increase the contribution of the successive factors as well. The centroid method of analysis as explained by Holzinger and Harman (1941) was programmed on an IBM

1620 Computer and the computations were carried out for three sets of genotypic and environmental correlation matrices (Tables 3, 4 and 5). The relative contributions of each factor and the percentage of total original communality accounted for by each of them are also given in Tables 3, 4 and 5.

Results

The centroid factor loadings obtained from the six correlation matrices are given in Tables 3, 4 and 5. Three factors were found to be adequate to explain the correlation matrices in this material since the factor matrices multiplied by their inverses were essentially equal to the original correlation matrices. The coefficients of the residual matrix were negligible in magnitude after eliminating the first three factors. The three factors can be considered to be essentially uncorrelated as evident from the size of the correlation coefficients (Tables 3, 4 and 5).

Table 3

Centroid factor matrix for 10 characters for genotypic and environmental correlation matrices in Sorghum, *1963 (Material A)*

Variable	Common factor coefficients						Communality			
	1		2		3		G		E	
	G	E	G	E	G	E	O	C	O	C
1	·252	·222	−·604	·002	−·152	−·415	·349	·451	·281	·221
2	−·359	·070	·448	·091	·106	·771	·231	·341	·335	·608
3	·913	·818	−·032	·192	·381	·162	·883	·980	·758	·732
4	−·034	·557	·449	−·702	−·663	·065	·435	·642	·787	·807
5	·884	·871	·240	·140	·118	·127	·885	·853	·758	·794
6	·345	·253	−·298	·137	·037	−·275	·345	·209	·228	·158
7	·456	·692	·657	·115	−·129	·000	·554	·656	·584	·492
8	·804	·734	·221	−·515	·218	−·065	·885	·743	·787	·808
9	·064	·196	−·606	·458	·222	·076	·349	·421	·284	·254
10	·265	·536	·419	·078	·073	−·446	·554	·251	·565	·492
Total							5·470	5·547	5·367	5. 366
Contribution of factor	2·856	3·200	1·938	1·070	·753	1·096				
% of total original communality	52·2	59·6	35·4	20·0	13·8	20·4	101·4		100·0	

Factor correlation matrix

	1		2	
	G	E	G	E
2	−·01	−·26		
3	·45	−·11	−·26	·06

G—Genotypic E—Environmental
O—Original C—Calculated

TABLE 4

Centroid factor matrix for 10 characters for genotypic and environmental correlation matrices in Sorghum, *1965 (Material B)*

Variable	Common factor coefficients						Communality			
	1		2		3		G		E	
	G	E	G	E	G	E	O	C	O	C
1.	..	·065	..	·237	..	−·175	..	..	·198	·091
2.	·867	−·059	−·162	·105	·229	·168	·897	·830	·058	·043
3.	·893	·207	−·166	·075	·266	−·488	·897	·896	·327	·287
4.	·792	·321	−·255	·181	·131	−·437	·773	·709	·327	·327
5.	·752	·508	−·043	·085	−·114	−·044	·652	·580	·287	·267
6.	−·607	·215	·203	·089	·252	·340	·523	·473	·586	·170
7.	·129	·365	·895	−·116	·226	·337	·966	·869	·287	·260
8.	·326	·091	·331	·056	·251	·058	·448	·279	·198	·015
9.	·330	·303	·591	−·710	−·478	·000	·778	·687	·596	·596
10.	·008	·146	·986	−·005	·260	·240	·966	1·040	·172	·079
Total							6·900	6·363	3·036	2·135
Contribution of factor	3·342	·713	2·394	·642	·627	·780				
% of total original communality	48·4	23·5	34·7	21·1	9·1	25·7		92·2		70·3

Factor correlation matrix

	1		2	
	G	E	G	E
2	−·61	−·27		
3	−·14	−·11	−·07	−·21

G – Genotypic E – Environmental
O – Original C – Calculated

TABLE 5

Centroid factor matrix for 12 characters for genotypic and environmental correlation matrices in Sorghum, *1965 (Material B)*

Variable	Common factor coefficients						Communality			
	1		2		3		G		E	
	G	E	G	E	G	E	O	C	O	C
1.	..	·028	..	·132	..	·198	..	..	·198	·057
2.	·848	−·041	·217	·009	·214	−·123	·897	·812	·058	·017
3.	·878	·147	·184	·230	·158	·060	·897	·830	·327	·078
4.	·812	·265	·018	·386	·178	·102	·773	·691	·327	·230
5.	·733	·347	·142	·058	·037	−·347	·652	·559	·287	·244
6.	−·642	·357	·105	·254	−·078	·172	·523	·429	·586	·222
7.	−·080	·311	·827	−·165	−·443	−·408	·966	·887	·287	·290
8.	·335	·104	·290	−·130	−·439	·113	·448	·389	·198	·040
9.	·052	·427	·690	−·573	·294	·096	·778	·565	·596	·520
10.	−·199	·124	·843	·023	−·524	−·330	·960	1·025	·172	·125
11.	−·120	·512	·727	−·516	·280	·198	·778	·621	·596	·568
12.	−·576	·522	·694	·291	·322	·267	·721	·971	·586	·428
Total							8·399	7·725	4·218	2·819
Contribution of factor	3·606	1·182	3·077	1·011	1·042	·626				
% of total original communality	42·9	28·0	36·6	24·0	12·4	14·8		91·9		66·8

Factor correlation matrix

	1		2	
	G	E	G	E
2	−·58	−·27		
3	·18	·23	−·22	·09

G–Genotypic E–Environmental
O–Original C–Calculated

Since factors contributing to yield and fitness only were included in this study, all of them were included in the factor study without setting a lower limit for the magnitude of their loadings on the factors to be included in the study as done by Sokal (1961). Arrow diagrams showing the important effects of factors on variables following the procedure adopted by Sokal (1961) for each case are presented in figures 1, 2 and 3.

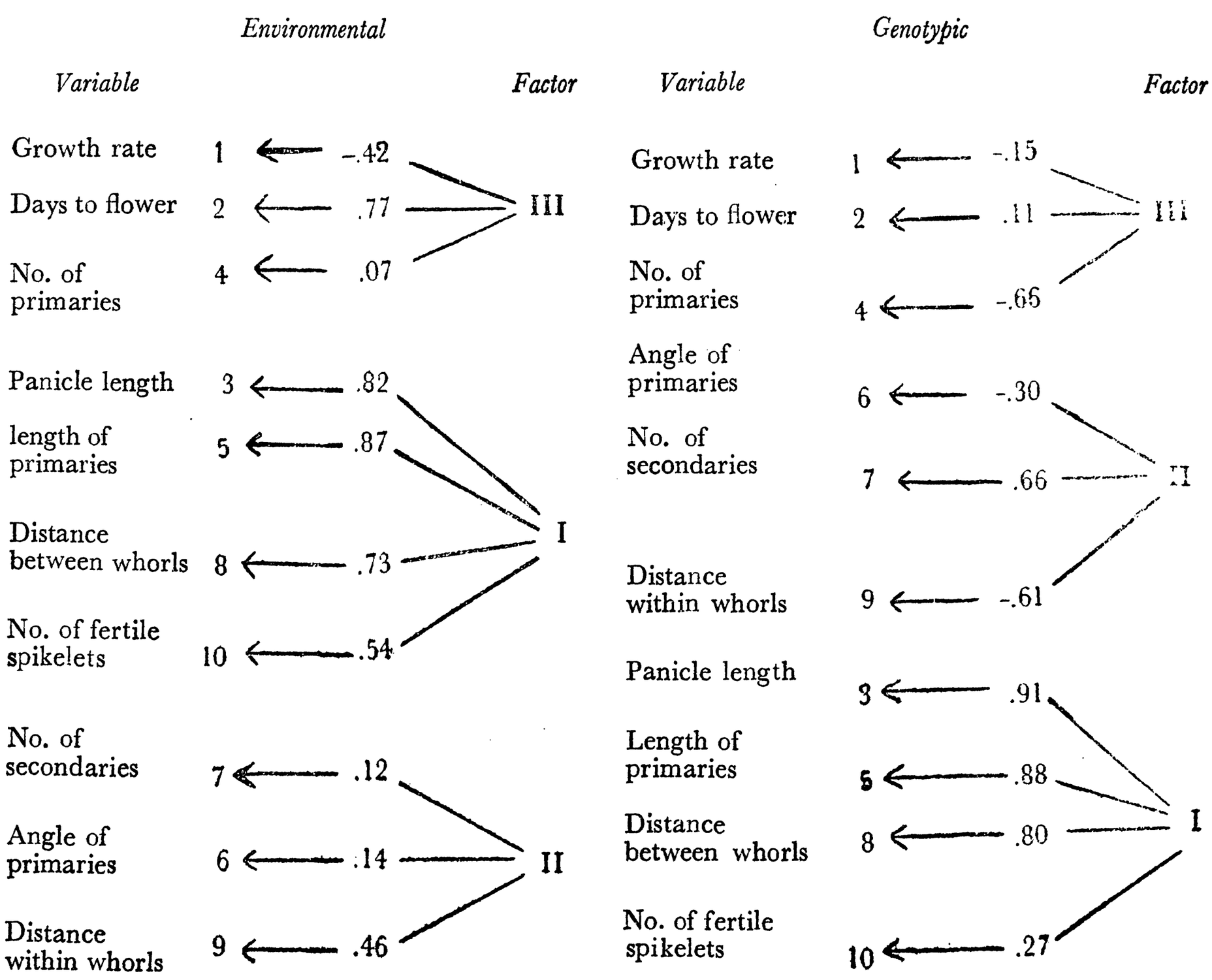

FIG. 1 Arrow diagram showing the effect of factors on ten variables in *Sorghum*, 1963

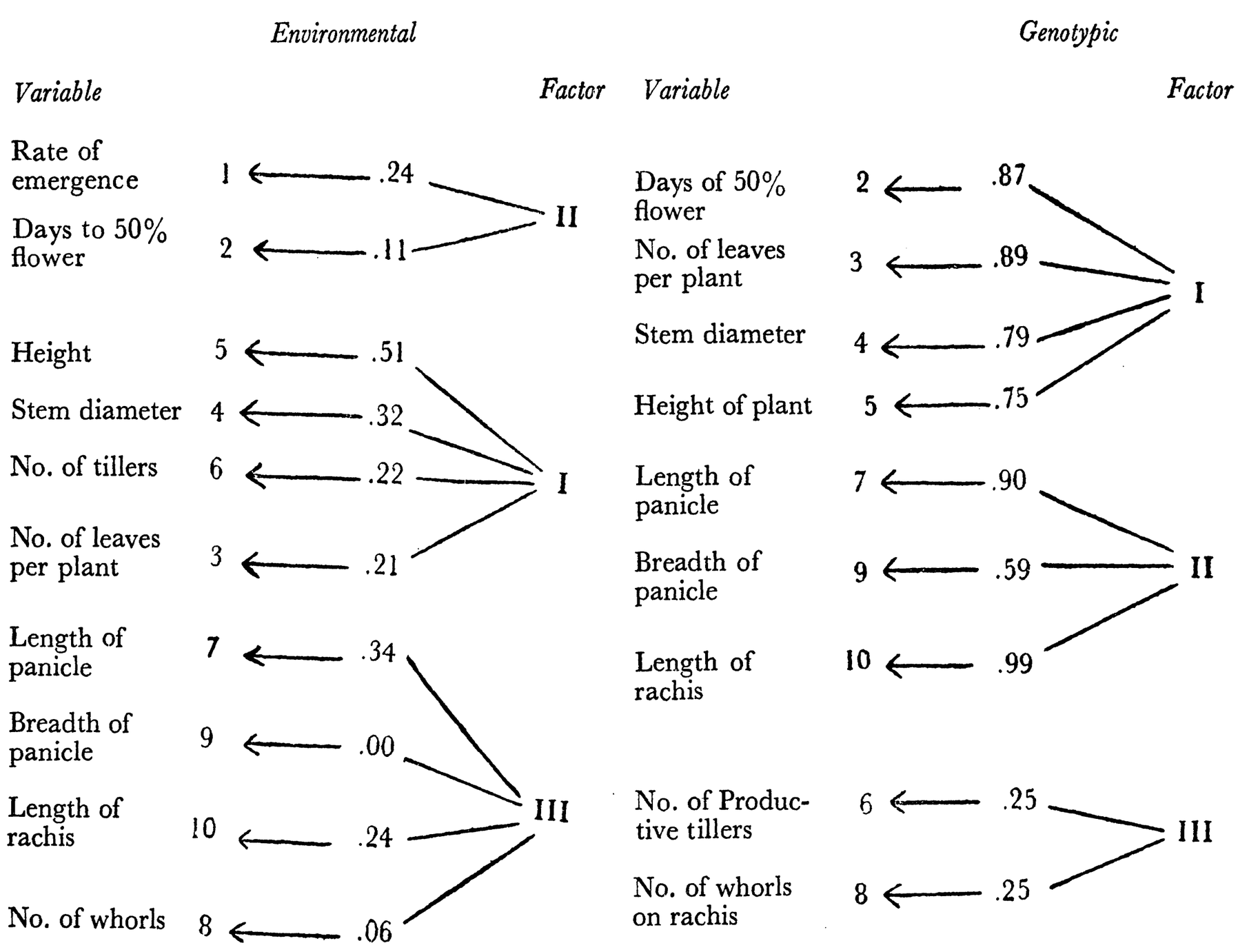

FIG. 2 Arrow diagram showing the effect of factors on ten variables in Sorghum, 1965

In material A, representing most of the species of the genus *Sorghum*, the ten characters could be considered to constitute three factors, these factors being the same for both genotypic and environmental correlations matrices. Growth rate, days to flower and number of primaries constituted one factor which could be considered as growth factor. Panicle length, length of primaries, distance between whorls and number of fertile spikelets which would determine the reproductive capacity formed another factor which was termed as reproductive factor. The other factor consisted of angle of primaries, number of secondaries and distance within whorls, which would determine the shape of the panicle and was, therefore, designated as panicle shape factor. It was significant that the same three factors were obtained for both the genotypic and environmental correlation matrices indicating the proper choice of the characters and their consistency of importance in different environments. The loading on factor I for number of secondaries in the case of environmental correlation matrix was more than that on factor II (Fig. 1). Similar was the case for growth rate and days to flower in the case of genotypic correlation matrix which were having higher loadings in factor II than in III in which they were included.

In the case of material B, consisting of 80 desirable breeding stocks from a world collection of *Sorghum*, the same factors did not explain genotypic and environmental inter-correlations. The composition of the factors also was different

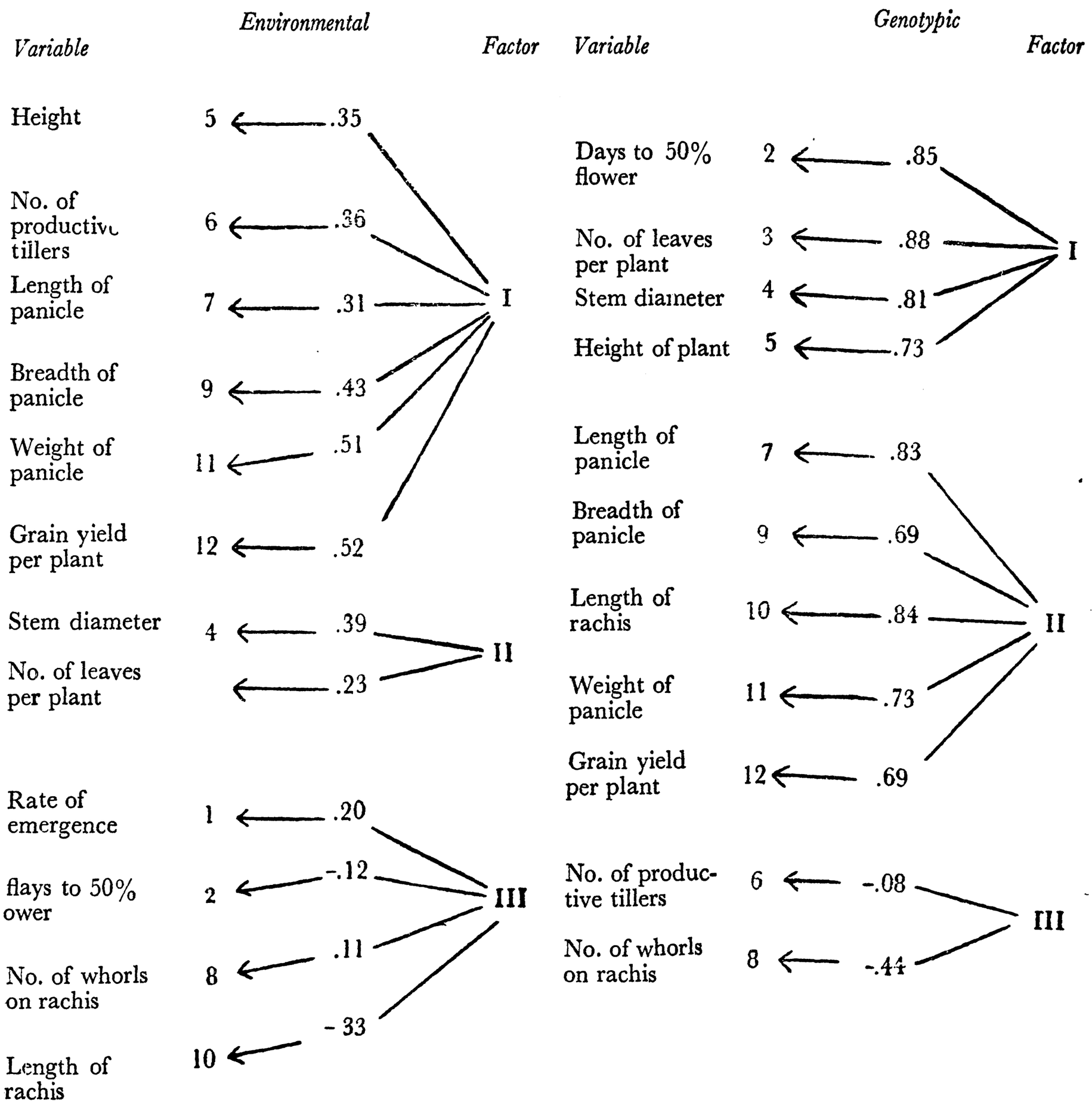

FIG. 3 Arrow diagram showing the effect of factors on 12 variables in *Sorghum*, 1965

from that of A. The genotypic correlation matrix was explained by three factors. The first factor consisted of days to 50 per cent. flower, number of leaves per plant, stem diameter and height of plant. The second factor included length of panicle, breadth of panicle, length of rachis and weight of panicle and grain yield per plant when the last two characters were included in the study while the third one was formed by number of productive tillers and number of whorls on rachis. It was interesting to note that the first and second factor appeared to be correlated ($r=-\cdot58$ and $-\cdot61$ when the yield characters were and were not included respectively).

On the other hand, factors obtained from the environmental correlation matrix excluding weight of panicle and grain yield per plant would appear to be more meaningful in this material. Inclusion of yield as a variable tended to

vitiate the picture, since it was not directly related to genetic diversity (Fig. 2). Factor 1, consisting of height, stem diameter, number of tillers and number of leaves per plant, was important for productivity and greater photosynthetic capacity, while factor 2 comprising of rate of emergence and days to 50 per cent. flower was an important factor for growth and therefore for adaptation. The third factor included length of panicle, breadth of panicle, length of rachis and number of whorls which would determine panicle shape was important for yield. It was observed that the length and breadth of panicle had higher loadings in factor I than factor III in the case of environmental correlation matrix.

A comparison of the contribution of the three factors to the total communality showed that the three factors accounted for 100 per cent. of the total communality in the case of genotypic and environmental correlation matrices in material A while the three factors accounted for about 92 per cent. in the case of genotypic and about 70 per cent. in the case of environmental correlation matrices irrespective of the inclusion of yield as a variable in material B.

Discussion

This study was initiated with a preliminary knowledge about the characters included. The characters chosen were important contributors to yield and fitness. Material A had already been subjected to studies on divergence and the characters chosen were found to be appropriate and adequate (Chandrasekeriah, 1964). Hence, the main aim in subjecting the same material to factor analysis was to find out whether this method would be able to provide fewer meaningful factors responsible for differentiation among species or populations in the genus *Sorghum*.

The study revealed the utility of factor analysis to provide fewer stable factors to delineate the divergent populations. It was possible to extract only three factors in this material in all the cases considered since the coefficients in the residual matrix were too low to allow extraction of more factors. It was found that the number of factors that should be extracted would depend on the material taken for investigation as indicated by Cattell (1965). The rule that $(p+k)$ should be less than $(p-k)^2$ where p is the number of characters and k the number of factors appears to be useful as indicated by Lawley and Maxwell (1963).

Divergent opinion exists among workers regarding the correct method of estimation of communalities. In fact, various methods had been tried by many workers in different fields in this connection. The results of the present investigation would indicate that the estimated communalities were adequate for drawing conclusions in material A where the three factors together accounted for 100 percent of the total communality in the case of genotypic as well as environmental correlation matrices(Table 3). However, in material B, the three factors contributed 92 per cent. to the total communality in the genotypic and 70 percent in the environmental correlation matrix, probably due to the fact that the material B

consisted of highly selected populations. Moreover, the centroid method of factor analysis would appear to be adequate for biological investigations of this nature as indicated by Holzinger and Harman (1941) and Thurstone (1947). It is, however, proposed to examine the nature of factors obtained by other methods of factor analysis in the same material.

Even with a smaller number of variables included in this study, factor analysis was potent enough to isolate the different factors responsible for differentiation. It was also significant that in material A, the same three factors relating to growth cycle and panicle were obtained from both the genotypic and environmental correlation matrices. While the principal component analysis indicated the adequacy of the first two canonical vectors (Chandrasekariah, Murty and Arunachalam, 1966), factor analysis revealed the adequacy of the first three factors for differentiation.

The stability of the factors was very high in the material representing the whole genus rather than in the material consisting of a number of selected populations. It is probable that stability of the factors from genotypic and environmental correlation matrices may be achieved by including a larger sample of the world collection.

The characters, weight of panicle and grain yield per plant appear to fit in the factor relating to panicle characteristics. Omission of these two yield characters from this factor enabled a better explanation of the correlation matrices by the factors.

The difference between the materials A and B in the composition of the factors was of interest. As stated earlier, the former represented the spectrum of diversity in the whole genus of *Sorghum* while the latter was limited to those highly favoured under domestication by man. There are several instances where human and natural selection operate in opposite directions. Therefore, the constellation of the selected characters and the correlations between them were substantially modified in B as compared to A. Moreover, the genetic correlations were highly skewed in B due to directional selection by man. A comparison of the factor loadings from genetic and environmental correlation matrices revealed substantial changes in the sizes of the loadings although the composition of variables in the factors remained essentially the same in A. This was to be expected since the genotypic variance-covariance matrix need not necessarily be an estimate of the parameter of a multivariate normal distribution. Therefore, it appears to be appropriate to use the common dispersion matrix represented by the environmental correlation matrix for factor analysis. The first, second and third factors in A are growth, reproductive and panicle shape factors respectively (Fig. 1). All these three constituted essentially causative influences of divergence under natural selection.

The factors in B although different in composition from those of A were similar in their function. However, factor I determined the productivity under cultivation since all its components were known to have been highly selected by man. Similar was the situation for the variables in factor III due to intense

selection for increased grain number per panicle. Factor II was important for adaptation.

The inclusion of yield per plant and panicle weight upset the loadings in each factor indicating that they were not important as such in divergence since most of the components of yield were already included as other variables.

The results confirmed that the pattern of divergence under natural selection was quite distinct from that under human selection in the genus *Sorghum*. Early growth and reproductive capacity appear to be the major causative influences for divergence rather than simple morphological features such as grain colour, glume colour and endosperm type which are of local importance even under selection by man. The present study provided a useful supplement to our earlier studies in this genus by generalised distance and principal component analysis.

Summary

The pattern of diversity in the genus *Sorghum* was analysed using the centroid method of factor analysis based on 10 to 12 characters in two groups of populations. One of them is representative of the spectrum of variation in the genus and the other comprised of high yielding grain types of Eu-sorghum.

Three factors were found to be adequate to account for most of the intercorrelations in both the genotypic and environmental correlation matrices.

The factor loadings on the variables were different in the two groups of populations indicating distinctly diverse causal differences under natural and human selection.

While the loadings were similar for genotypic and environmental correlations in natural populations, the differences were marked in the second group of selected populations. The data indicated that the environmental correlation matrix is appropriate for factor analysis in this material. The study revealed that the inclusion of variables influencing yield is more appropriate than yield itself in multivariate analysis.

The three factors in the group of natural populations were found to be growth, reproductive and panicle shape factors all known to be important components of fitness.

The results of factor analysis provided supplementary information on the diversity in this genus not available from the analysis of principal components and generalised distance and on the adequacy of the centroid method in biological investigations.

REFERENCES

Cattell, R.B. (1965). Factor Analysis: An introduction to essentials I. The purpose and underlying models. *Biometrics*, **21:** 190–215.

———— (1965). Factor Analysis : An introduction to essentials II. The role of factor analysis in research. *Biometrics*, **21:** 405–35.

Chandrasekariah, S. R. (1964). Assessment of divergence by multivariate analysis in the genus Sorghum. Ph. D. Thesis, I.A.R.I., Delhi.

Chandrasekariah, S.R., Murty, B.R. and Arunachalam, V. (1967). Multivariate analysis of divergence in Eu-sorghums. Proc. Nat. Instt. Sci. **B** (In Press).
Holzinger, K. J. and Harman, H.H. (1941). *Factor Analysis*, University of Chicago Press, Chicago, U.S.A.
Lawley, D.N. and Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*, Butterworths, London.
Maxwell, A.E. (1961). Recent trends in factor analysis. *J. Roy. Statis. Soc., A.*, **124:** 49–59.
Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya*, **26**A: 329–58.
Snowden, J.D. (1936). *The Cultivated Races of Sorghum. Adlard and Son Ltd.*, London.
Sokal, R.R. and Daly, H.V. (1961). An application of factor analysis to insect behaviour. *Univ. Kansas Sci. Bull.*, **42:** 1067–97.
Thurstone, L.L. (1947). *Multiple factor analysis.* University of Chicago Press, Chicago, U.S.A.