Principal Component Analysis : Its Application to Classfication and Selection in Relation to Maize Breeding

By NOBORU MOCHIZUKI

Researcher, 2nd Laboratory, Division of Genetics, Department of Physiology and Genetics National Institute of Agricultural Sciences

Maize breeding starts from the collection of adapted local strains and the introduction of exotic germplasm in regions with similar climates, considering genetic diversity. Classification of the strains based on their origin and the evaluation of their characteristics in introduction fields is a basis to determine and select superior materials for breeding hybrids as well as the studies on variability and differentiation of races.

Principal component analysis, a method of multivariate statistical analysis, was successfully applied to the classification of Japanese local strains of maize and to the selection of high combinable strains from them for use in hybrids and synthetic varieties.

Principal component analysis of the characteristics of local flint in Japan

In the early days of hybrid maize breeding program in Japan, it was made clear that the hybrids between Japanese local flint and U.S. dent showed significant heterosis with respect to yield and adaptability. So since 1954 effort was concentrated in the collection of Caribbean flint local varieties mainly distributed in the mountainous areas of central and southern Japan. Up to the present approximately 600 strains were collected and evaluated.

These Caribbean flints were originally introduced by the Portuguese about 400 years ago. Since then farmers had planted them for staple food and livestock feed, resulting in high productivity and varietal differentiation.

The source material in this study was the data on the characteristics of representative local strains, open-pollinated varieties, of the Caribbean flint in Japan. The strains including 24 from each of Fuji, Shikoku, and Kyushu were observed at Hiratsuka, Kanagawa Prefecture in 1958. The origin of the materials is given in Fig. 1.

The objective of the principal component analysis is to produce a linearly transformed set of new variates, X_1, X_2, . . . , X_p, called principal components of original variates, x_1, x_2 . . . . x_p. Principal components are mutually independent and thus can be considered separately.

Since variances associated with the principal components are in decreasing order, it is possible that only a few variates are needed to summarize the whole of variability and covariability of the original variates, x's. Therefore, the principal component analysis is called a method of "parsimonious summarization of a mass of observation" (Seal, 1964).

Twelve characters out of 65 observed ones were selected for the analysis. The characters were silking date, stalk length, leaf length, leaf width, number of leaves, tassel length, ear length, ear diameter, ear weight, number of ears, 100 kernels weight, and grain yield.

The correlation coefficients between these characters were calculated (Table 1), following the principal component analysis. The resulting in eigen values which were variances

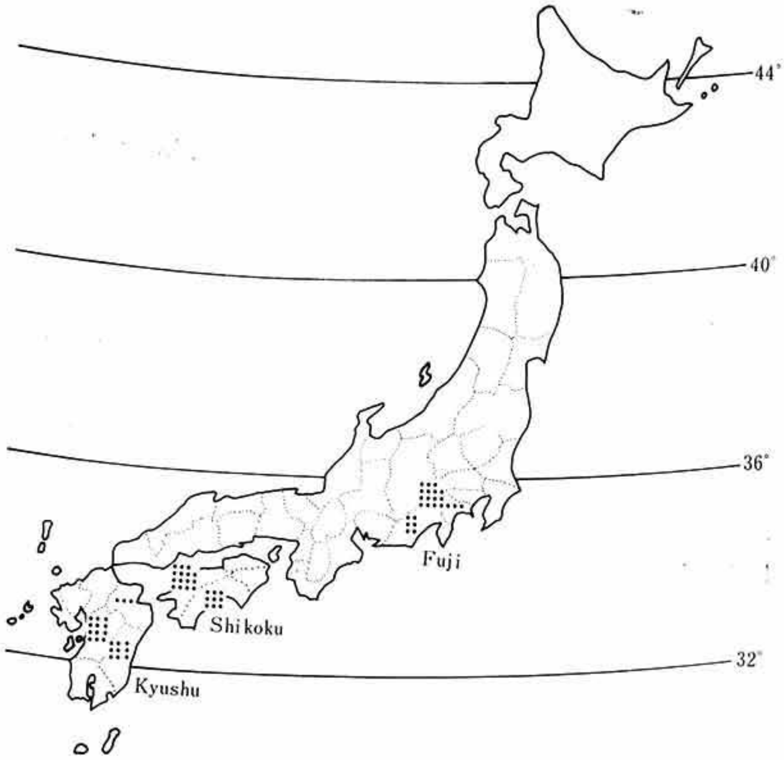

Fig. 1. Origin of 72 representative local strains of Caribbean flint in Japan.

Table 1. Correlation matrix of 12 characters in representative 72 Caribbean flint local strains in Japan

Character		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Silking date	X_1	1.000	.668	.456	.380	.884	−.297	.449	.565	.647	.430	−.171	.710
Stalk length	X_2		1.000	.216	.391	.818	−.106	.453	.371	.570	.304	.003	.593
Leaf length	X_3			1.000	.313	.372	−.005	.230	.427	.449	.392	−.129	.532
Leaf width	X_4				1.000	.394	−.146	.355	.353	.570	.126	.208	.486
Number of leaves	X_5					1.000	−.387	.366	.584	.657	.476	−.171	.741
Tassel length	X_6						1.000	.163	−.273	−.058	−.253	.338	−.157
Ear length	X_7							1.000	.015	.482	.183	.241	.439
Ear diameter	X_8								1.000	.803	.124	.156	.619
Ear weight	X_9									1.000	.178	.311	.798
Number of ears	X_{10}										1.000	−.370	.676
100 kernels weight	X_{11}											1.000	−.008
Grain yield	X_{12}												1.000

associated with the transformed variates indicated that 46, 16, 10, and eight percent of the total variation were accounted for by the first, second, third, and fourth principal components, respectively (Table 2).

Hence, nearly 80 percent of the total variation could be expressed. This means that the first four principal components were needed

Table 2. Eigen value (λ_k) and associated eigen vector ($l_{k1}, l_{k2}\cdots, l_{kp}$) obtained from principal component analysis of the 12×12 correlation matrix

Eigen value		X_1	X_2	X_3	X_4
λ_k		5.473	1.883	1.207	1.006
$\lambda_{k/p}$ (%)		45.6	15.7	10.1	8.4
$\sum_{l=1}^{k}\lambda_l$		5.473	7.356	8.563	9.569
$\sum_{l=1}^{k}\lambda_{l/p}$ (%)		45.6	61.3	71.4	79.7
Eigen vector		l_{1i}	l_{2i}	l_{3i}	l_{4i}
Silking date	x_1	.374	-.113	.031	-.140
Stalk length	x_2	.322	.035	.154	-.416
Leaf length	x_3	.246	-.052	.090	.674
Leaf width	x_4	.250	.228	-.089	-.041
Number of leaves	x_5	.385	-.150	-.025	-.264
Tassel length	x_6	-.117	.429	.434	.318
Ear length	x_7	.217	.294	.534	-.210
Ear diameter	x_8	.301	.082	-.548	.208
Ear weight	x_9	.365	.285	-.183	.085
Number of ears	x_{10}	.224	-.386	.352	.225
100 kernels weight	x_{11}	-.005	.633	-.140	.034
Grain yield	x_{12}	.388	-.022	.091	.184

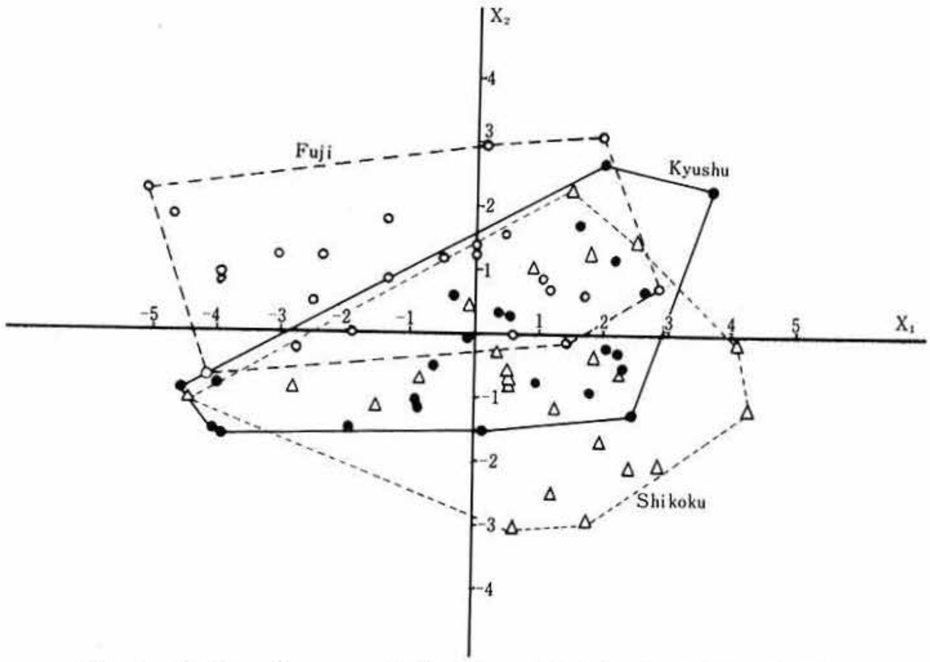

Fig. 2. Scatter diagram of Caribbean flint local strains collected from Fuji, Shikoku, and Kyushu projected in the (X_1–X_2) plane.
X_1: First principal component.
X_2: Second principal component.

to summarize the whole of the variability and covariability in the characteristics of the local strains. The eigen vectors corresponding to the respective eigen values are presented in the lower half of the table.

The scatter diagram (Fig. 2) of the 72 strains on the axes of the first and second principal components suggested that regional differentiation existed in the characteristics of the flint local strains.

Another point of interest is to interpret the principal components in biological terms.

The principal components are linear transformation type as

$$X_k=\sum_{i=1}^{p} l_{ki} x_i$$

where X_k represents the kth principal component, l_{ki} ith element of eigen vector corresponding to the kth principal component, x_i ith character. For example

$$X_1=0.374x_1+0.322x_2+0.246x_3+0.250x_4 +0.385x_5-0.117x_6+0.217x_7+0.301x_8 +0.365x_9+0.224x_{10}-0.005x_{11}+0.388x_{12}$$

Thus, each of the principal components could be interpreted as compound characters or plant types respectively, which were mutually uncorrelated. The plant types discriminated by negative or positive directions on the axis of the respective principal components are given in short as follows:

First principal component: early maturity, short plant height, small ear, and low yield vs. late maturity, high plant height, large ear, and high yield.

Second principal component: single large ear, large kernel size vs. prolific small ear, small kernel size.

Third principal component: long slender ear vs. short conical ear. Fourth principal component: short plant height and long leaves vs. high plant height and short leaves.

The biological meaning of the first principal component appeared to correspond to the general "size" of plant in relation to the growing period, the second distribution of photosynthetic products in plant, and the third morphological variation in ear shape. Both, and also the fourth one were thus indicator of "shape".

Classification of Caribbean flint in Japan

So as to classify the strains into strain groups or varieties having similar characteristics, the square distance between the 72 strains in the four dimensional space was calculated from the first four principal component's score of the strains. The smaller the squared distance was between strains, the more similar the characteristics of the strains were expected to be. So the strains among which the squared distances were very small were grouped as a variety.

The criterion of the grouping was that the average distance within a variety was always smaller than the ones among varieties. In consequence, the 72 strains were classified into 14 varieties. Furthermore, by the same way, the varieties were classified into four varietal

Table 3. Classification of representative Caribbean flint local strains in Japan

Varietal Group	Variety	Typical local variety
A	V_1	Irareko (S), Hachiretsu-wase (K)
	V_2	Narusawa (F), Hirano (F)
B	V_3	Kamigane (F), Doshi (F), Sengoku (S), Abetto (S), Nakadama (K)
	V_4	Yusuhara (S), Odecchi-Kuju (K)
	V_5	Okuzuru (K)
	V_6	Iwama (F)
	V_7	Gojo (S)
	V_8	Shinboso (K)
	V_9	Kowase (S)
	V_{10}	Akiyama (F)
	V_{11}	Suginazawa (F)
	V_{12}	Suginazawa (F)
C	V_{13}	Yusuhara (S)
D	V_{14}	Wada (S)

Note: Letter in parenthesis indicates the origin of the local variety: F; Fuji, S; Shikoku, K; Kyushu.

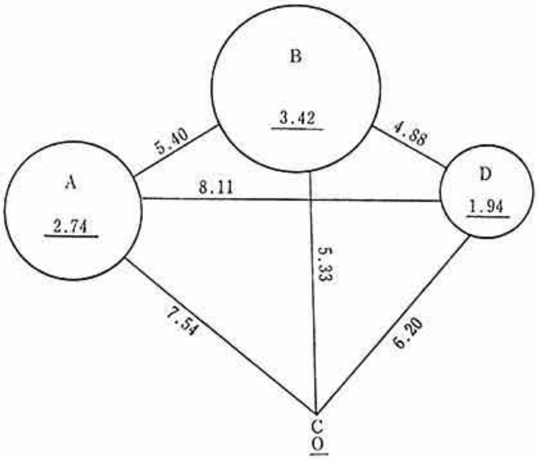

Fig. 3. Cluster representing varietal groups and their interrelationships in Caribbean flint local strains. Figure with underline in diagram indicates average distance within a varietal group, and figure without underline average distance between varietal groups.

groups such as A, B, C, and D.

The new statistical classification using the principal component analysis and the distance method is presented in Table 3. From this table it was made clear that V_3 was the variety with wide adaptability, distributing in Fuji, Shikoku, and Kyushu: however, V_{14} was the one variety which belonged to the varietal group D and distributed only in Shikoku. The interrelationship of the varietal groups in the four dimensional space is shown in Fig. 3. The figure in the diagram indicates the average distance between the varietal groups and within one.

The characteristics of the varietal groups may be summed up as follows:

Varietal group A: early maturity, short plant height, few leaves, single short and slender ear, medium to large kernel, and low yield.

Varietal group B: medium to late maturity, medium to high plant height, medium to many leaves, medium to long ear, medium ear diameter, and medium to high yield.

Varietal group C: medium maturity, medium plant height, narrow but many leaves, prolific ears, small kernel, and high yield.

Varietal group D: late maturity, medium plant height, long and many leaves, short but big conical ear, medium kernel size, and high yield.

Furthermore, most of the representative local varieties, which had been evaluated as superior breeding stocks in use for hybrids and actually had been utilized as parental sources of the recommended hybrids in Japan, belonged to the varietal group B in this classification. These local varieties were Ehime-otomorokoshi Nos. 1 and 3, Okuzuru-wase, Kagawa-zairai, Gojo, Akatokibi, Odecchi, Abetto, and Hakushoku-zairai. This fact suggested that the statistical classification based on the principal component analysis might be of significance for the selection of superior breeding materials.

Correlation between principal component and combining ability

For the purpose of verifying the applicability of the principal component analysis to preliminary selection of breeding materials from the many strains, further study was carried out on the relation between the principal component and combining ability. The material in this particular study consisted of two sets of data. One was on the adaptability trial of the 72 local strains including all of the ones analyzed before they were observed at Kuma, Ehime in 1958. The other was on the combining ability trial of the 24 strains from Fuji and 24 from Shikoku at the same location in 1960 and 1961, respectively.

Combining ability of these strains was tested in top cross trial with four U.S. dent testers, three inbred lines and a single cross.

On the adaptability trial data ten agronomic characters were selected, following the principal component analysis and classification of the strains. General and specific combining ability was estimated on combining ability trial by the procedures suggested by Federer and Sprague (1947) and Plasted *et al.* (1962).

Table 4. Correlation coefficient between principal component and combining ability

General and specific combining ability	Principal component			
	X_1	X_2	X_3	X_4
Strains of Fuji				
General combining ability	.919**	.104	-.302	.235
Deviation of specific combining ability in a strain	.495*	.117	-.143	.314
Strains of Shikoku				
General combining ability	.645**	.216	-.207	-.353
Deviation of specific combining ability in a strain	-.020	.210	.329	.056

Note: * and ** indicates statistical significance at 5% and 1%, respectively.

The result may be summarized as follows: High correlation coefficients between the first principal component's scores and the estimates of the general combining ability of strains were obtained, i.e. $r=0.919^{**}$ and 0.645^{**} in the strains from Fuji and Shikoku (Table 4). These correlation coefficients were higher than that between the grain yield of strain *per se* and the estimates of the general combining ability, i.e. $r=0.762^{**}$ and 0.502^{**}, respectively.

Regarding the other principal components no correlations were found. Also no or low correlation was observed between all the four principal components and the standard deviations of the specific combining ability in a strain.

The 72 strains were grouped into 13 varieties which were further classified into six varietal groups. Most of the strains showing high general combining ability were included in a particular varietal group, the B group in the present study. These results indicated that preliminary selection of breeding materials without testing procedure of the combining ability were possible by the application of the principal component analysis to the data on characteristics of strains obtained in the introduction field.

Thus, it was concluded that classification of many strains collected and introduced and preliminary selection of superior breeding materials for use in hybrids and synthetic varieties could be achieved by application of the principal component analysis.

References

1) Federer, W. T. and Sprague, G. F.: A Comparison of variance components in corn yield trials: Error, Tester × Line, and line components in top cross experiments. *Jour. Amer. Soc. Agron.* **39**, 453–463 (1947).
2) Kendall, M. G.: A Course in Multivariate Analysis. Charles Griffin & Co. Ltd. London. 185 (1957).
3) Mochizuki, N. and Okuno, T.: Classification of maize lines collected from Shikoku, Japan, and selection of breeding materials by the application of the principal component analysis. *Japan Jour. Breeding.* **17**, 39–47 (1967).
4) Mochizuki, N.: Classification of maize lines and selection of breeding materials by the application of principal component analysis. *Bull. Nat'l. Inst. Agric. Sci.* Series D **19**, 85–149 (1968).
5) Plasted, R. L., Sanford, L., Federer, W. F., Kehr, A. E. and Peterson, L. C.: Specific and general combining ability for yield in potatoes. *Amer. Potato Jour.* **39**, 185–197 (1962).
6) Seal, H.: Multivariate Statistical Analysis for Biologists. Methuen & Co. Ltd. London. 209 (1964).