Jour. Ind. Soc. Ag. Statistics
48(3), 1996 : 339–345

An Alternative Approach for the Analysis of Data of Long Term Experiments

P.V. Prabhakaran and Rani John
Kerala Agricultural University, Thrissur - 680 654

(Received : September 1990)

SUMMARY

A new method was suggested for the analysis of data of repetitive trials with dependent sets of observations. The method consists in extracting the first Principal component from the original Nr × p matrix of observations where N is the number of treatments, r is the number of replications and p is the number of years. As the Principal component is the best linear combination of the 'p' yearly responses the assumption of independence of error terms seems to be logically sound. In situations where two or more components have to be retained, multivariate analysis of variance of the transformed component scores may be attempted, the transformation being the division of the scores by the square root of the respective eigen values. The method was applied to the data generated from the permanent manurial trial on rice at Pattambi and the results indicated that it is slightly more efficient than the usual split-plot analysis and the analysis of groups of experiments.

Key words: Long term experiments, Joint statistical analysis, Treatment effect, Independence of error terms, Principal component analysis, Index score, Multivariate analysis.

1. Introduction

In large scale experimental programmes it is necessary to repeat a trial of a set of treatments at a number of places and in a number of years in order to know the susceptibility of the treatment effects to place and climatic variations. The usual practice followed in such repetitive trials is to perform a joint statistical analyses of data by using the analysis of variance technique as applied to groups of experiments on the assumption of independence of error terms which will not be usually valid. Further in such types of analysis no general test appears to be available for overall treatment comparisons when error variances are heterogeneous and interaction effect is absent. Another possibility in dealing with such experiments is to consider them as special cases of a split-plot arrangement with years or seasons as subplots, within each treatment mainplot. But, here also the assumption of independence of error terms does not seem to be wholly valid. It is therefore necessary to find an alternative to the analysis of groups of experiments and split- plot analysis so as to draw fairly accurate inferences regarding the suitability of treatments. The present study is aimed at examining the utility of the technique of principal component analysis in the interpretation of data from long term trials with dependent sets of observations. Danford *et al.* [2] and Cole and Grizzle [1] have applied multivariate techniques such as Hotteling's T^2 and likelihood ratio criterion for the analysis of data from long term trials and the results were promising.

2. *Materials and Methods*

Consider the random variables, $x_1, x_2, \ldots\ldots, x_p$ which have a multivariate distribution with mean vector μ and correlation matrix Σ. Assume that the elements of μ and Σ are finite. Let the rank of Σ be p and the 'p' characteristic roots be $\lambda_1, \lambda_2, \ldots, \lambda_p$ such that $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Let there be N treatments repeated over p years. The observations (X_{ij}) can be written in the form of $N \times p$ data matrix.

Transform X_{ij} to standard score Z_{ij} as

$$Z_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j}, \qquad (i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, p) \qquad (1)$$

where $\overline{X}_j$ and S_j are respectively the mean and standard deviation of X_j. The covariance matrix of $Z = (Z_{ij})$ will be the correlation matrix of the original data matrix and will be of order $p \times p$.

The first principal component of the observations Y_1 is that linear compound defined by,

$Y_1 = a_{11}Z_1 + a_{21}Z_2 + \ldots + a_{p1} Z_p = a_1' Z$ such that $a_1' a_1 = 1$ and variance of y_1 is maximum. The coefficients of this linear equation must satisfy the p simultaneous linear equations, $(\Sigma - \lambda_1 I)\, a_1 = 0$. The value of λ_1 must be so chosen as to make $|\Sigma - \lambda_1 I| = 0$. λ_1 is thus a characteristic root of the correlation matrix and a_1 is its associated characteristic vector. Similarly all other characteristic roots and characteristic vectors can be found out so that $\lambda_1 + \lambda_2 + \ldots + \lambda_p = \text{trace}\, \Sigma = p$.

The first principal component serves as that linear combination of years which explains maximum variation among treatments. This is simply a weighted index of seasonal components, the weights being the coefficients in the associated eigen vector. The process provides a unique value for each treatment which is obtained by multiplying the transformed matrix 'Z' with the eigen vector a_1. This value of the derived composite variable known as the index value acts as an index of performance of specific treatments in relation to others and thus helps in the discrimination between treatments. The treatments are then ranked on the basis of the indices and the best treatment is recommended for adoption.

But the method described above fails to provide a statistical test of significance. A more general approach is to derive the principal components from the original Nr × p matrix of observations where r is the number of replications for each treatment. Standardised values are then obtained by applying the transformation described in (1).

Principal components are extracted as, $$Y_m = \sum_{j=1}^{p} a_{mj} Z_{ij} \qquad (2)$$

where Y_m = m^{th} Principal component (m = 1, 2, 3, ..., p)

a_{mj} = The coefficient of the j^{th} variable (year) for the m^{th} component

Z_{ij} = Standardised value of X_{ij}

The percentage variability attributed by the j^{th} component is $\lambda_{j/p}$. In case the first principal component alone could explain a sizable amount of variability, say, 75%, the other components need not be considered. The component scores for the selected components are obtained by multiplying the Nr × p matrix of standardised values with the eigen vector of order p. Data of the two way table involving the index scores of the N treatments in r replications can be analysed as in a randomised block design.

If the percentage variability explained by the first principal component was relatively small two or more components may have to be selected for the description of the data. In such cases multivariate analysis using the transformed scores on the selected components may be attempted. The relevant transformation for the stabilisation of variance being, $Z_{ijk} = Z_{ij}/\sqrt{\lambda_k}$ where Z_{ijk} is the component score on the k^{th} component corresponding to ith treatment in j^{th} block and λ_k is the latent root of the k^{th} component.

3. *Empirical Evaluation*

The method described above was applied to the analysis of the twelve year yield data of the permanent manurial trial on rice collected from the Regional Agricultural Research Station, Pattambi, Kerala. The experiment was laid out in a 4 replicate randomised block design with 8 treatments. The treatments were different combinations of organic and inorganic manures.

The eigen values and eigen vectors generated from the correlation matrix of the observations and the percentage variation explained by the component vectors are given in Table 1. Since the first principal component explained more than 75 per cent of total variation the other components were not considered for the analysis. The transformed matrix Z was then multiplied by the eigen vector corresponding to the largest eigen value and the index values (component scores) for each treatment was obtained. The index values of the treatments were as follows.

Treat-ments	T1	T2	T3	T4	T5	T6	T7	T8
Index Values	3.5269	–2.1098	2.4228	–4.1254	3.2945	–2.1344	1.5468	–2.410

In the general case, the original 32 × 12 matrix of observations was transformed into a matrix of standardised values. Eigen values and corresponding eigen vectors were generated from this matrix. Then, by multiplying the 32 × 12 matrix of standardised values with the largest eigen vector of order 12 an index score matrix of order 32 × 1 was obtained which was rearranged in the form of a two-way table of treatments and replications. The data were further analysed as in a randomised block design and the treatment effect was found to be significant. Comparisons were also made between pairs of treatments using the calculated critical difference. The result obtained is as given below.

$$T_5 \quad T_1 \quad T_3 \quad T_7 \quad T_6 \quad T_8 \quad T_2 \quad T_4$$

Analysis of groups of experiments and split-plot analysis also gave identical results. Since all these three methods make use of 'F' test for testing the significance of treatment effects the relative efficiencies of the methods can be empirically compared on the basis of the relative magnitude of the relevant 'F' ratios. The 'F' values for testing the overall treatment effects as obtained from the three methods were as follows.

Methods	*'F' values*
Groups of experiments	24.49**
Split-plot analysis	27.83**
Principal component analysis	27.89**

Principal component analysis recorded the maximum 'F' value for detecting the real treatment effect. Percentage variation explained by overall treatment differences in the three methods of analysis was found to be 20.63

Table 1. Eigen values and corresponding eigen vectors

	Eigen vectors											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
	0.0957	−0.7015	−0.3475	−0.3650	0.1374	0.2225	−0.0855	−0.0939	0.1629	0.0375	−0.1434	0.3343
	0.3080	−0.2601	−0.1533	−0.1635	−0.1940	−0.0339	0.2430	0.4469	−0.1487	−0.4506	0.1768	−0.4830
	0.2363	−0.3508	0.5399	0.3736	0.4224	0.1118	0.0539	−0.1636	−0.1549	−0.2324	0.2824	0.1079
	0.3081	0.0099	−0.2378	0.4015	−0.5768	0.1295	−0.2122	−0.3150	0.2478	−0.1825	0.2761	0.1501
	0.2827	−0.2928	0.3024	−0.0049	−0.4128	−0.5059	−0.0037	0.1269	−0.2363	0.4507	−0.1050	0.1742
	0.3272	−0.0077	0.1081	0.0220	−0.0681	0.5307	0.1617	−0.2609	−0.0462	0.4264	0.4264	−0.2864
	0.2999	0.1186	−0.4003	−0.0586	0.3245	−0.4475	0.3857	−0.4298	−0.0185	0.1543	0.2537	−0.0743
	0.3073	0.2269	−0.2665	0.1146	0.0956	0.1817	−0.1578	0.0079	−0.6546	−0.2349	−0.4140	0.2722
	0.2592	0.3001	0.3423	−0.7169	−0.1156	0.1246	−0.1347	−0.2149	−0.0909	−0.1426	0.2639	0.1572
	0.3126	0.2518	−0.0043	0.0825	0.0661	0.2530	0.4618	0.4849	0.3775	0.1667	0.0697	0.4549
	0.3239	0.0738	0.1997	−0.0059	0.1193	−0.3083	−0.0006	−0.0632	0.5356	−0.3276	−0.5126	−0.0888
	0.3227	0.0826	−0.1325	0.0366	0.3330	−0.0316	−0.6698	0.3334	0.1075	0.2983	0.2494	−0.1890
Eigen values	9.0592	1.5754	0.8021	0.2956	0.1319	0.0719	0.0369	.000077	.000019	−.000009	−.000057	−.000088
% variation explained by eigen vector	75.49	13.13	6.68	2.46	1.10	0.60	0.31	0.0006	0.0008	0.00006	0.0005	0.0007

Table 2. Two way table of data generated through principal component analysis

Treatments	*Replications*				*Total*
	R_1	R_2	R_3	R_4	
T_1	1.8505	3.1743	3.1647	2.9675	11.157
T_2	-2.4517	-2.4703	-2.5261	-0.7269	-8.175
T_3	2.0856	2.2002	1.3013	2.8334	8.4205
T_4	-4.7964	-2.5719	-3.3912	-3.5168	- 14.2763
T_5	3.1458	2.2371	3.7412	2.6884	11.8125
T_6	0.6423	-2.1477	-2.7397	-2.4507	-6.7021
T_7	0.7876	1.8573	1.3394	1.5106	5.4949
T_8	- 1.2876	- 2.2350	- 0.8989	- 3.3208	- 7.7423
Total	- 0.0239	0.0440	- 0.0093	- 0.0216	- 0.0108

in analysis of groups of experiments, 21.94 in split-plot analysis and 90.29 in principal component analysis. Thus principal component analysis yielded better predictability for the overall treatment comparisons than the other two methods. Therefore, principal component analysis may be preferred to the conventional methods for the analysis and interpretation of data from long term trials.

However, the method has the following limitations

1. The method is useful in case the variation explained by the first principal component is substantially large, preferably more than 75 per cent.
2. The interpretation of the analysed data through the present method has to be done only for the transformed variables.

ACKNOWLEDGEMENT

The authors are thankful to the referee for his useful comments.

REFERENCES

[1] Cole, J.W.C. and Grizzle, J.E., 1966. Application of multivariate analysis of variance to repeated measurements experiments. *Biometrics*, **22**, 810-828.

[2] Danford, M.B., Hughes, H.M. and Mc Nee, R.C., 1960. On the analysis of repeated measurements experiments. *Biometrics*, **16**, 547-564.

[3] Khosla, R.K., Rao, P.P. and Das, M.N., 1979. A note on the study of experimental error in groups of agricultural field experiments conducted in different years. *J. Indian Soc. Agric. Stat.* **31**, 65-68.

[4] Patterson, 1939. *Statistical Techniques in Agricultural Research*. Mac Graw Hill & Co., New York.

[5] Yates, F. and Cochran, W.G., 1938. The analysis of groups of experiments. *J. Agric. Sci.*, **28**, 556-580.