Euphytica **44**: 109–123, 1989.

An investigation of multi-attribute genotype response across environments using three-mode principal component analysis

P.M. Kroonenberg[1] and K.E. Basford[2]
Department of Psychology, University of Queensland, Australia; [1] present address: Department of Education, University of Leiden, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands; [2] present address: Department of Agriculture, University of Queensland, St. Lucia, Queensland 4067 Australia

Received 23 February 1988; accepted in revised form 28 September 1988

Key words: three-mode principal component analysis, soybean lines, ordination, multivariate analysis, genotype-environment interaction

Summary

The usefulness of three-mode principal component analysis to explore multi-attribute genotype-environment interaction is investigated. The technique provides a general description of the underlying patterns present in the data in terms of interactions of the three quantities (attributes, genotypes, and environments) involved. As an example, data from an Australian experiment on the breeding of soybean lines are treated in depth.

Introduction

The existence of significant genotype × environment interaction creates difficulty in genetic analysis in several ways, such as by confounding estimates of genetic parameters and statistics, and by complicating selection and testing strategies. Such interactions reflect differences in adaptation which may be exploited by selection and by adjustments to the test strategy. In this context, conflict inevitably exists between breeding for *broad adaptation* (minimizing interactions) and *specific adaptation* (emphasizing favourable interactions). However, any objective decision requires a full understanding of the nature of genotype × environment interactions. Further complications arise because commonly, breeders are interested in more than one attribute at a time. Selection indices (Smith, 1936; Manning, 1956) were an early attempt to combine multi-attribute information into a single variable for subsequent analysis.

In this paper a multivariate technique, Three-Mode Principal Component Analysis (TMPCA) is used to handle all genotypes, environments, and attributes simultaneously. The primary aim will be to demonstrate how the technique can give a general description of the main patterns present in the data in terms of interactions of the three quantities involved.

The technique will be illustrated with data on the adequacy of several lines of soybeans (*genotypes*) scored on several characteristics (*attributes*) at several locations measured in two consecutive years (*environments*). Previous analyses of these data have been published, notably using ordination and classification (Mungomery et al., 1974; Shorter et al., 1977), individual differences scaling (Basford, 1982), and three-way clustering (Basford & McLachlan, 1985). The adaptation of the genotypes is, therefore, well known so the use of this data set permits some judgement on the usefullness

of this method of analysis. The analyses reported here should be feasible for any genotype by environment by attribute data.

Experimental details

Mungomery et al. (1974) is the first published account of the experiment from which these data were collected. Fifty-eight soybean lines, whose origin and maturity details are shown in Table 1, were evaluated at four locations in south-eastern Queensland in 1970 and 1971. The first forty breeding lines were local selections obtained from crossing line 43 (Mamloxi) with line 41 (Avoyelles). As only a few of these were released as varieties (and so given a cultivar name) they will only be referred to in the subsequent text by line number. Lines 41 to 58 will be referred to by line number with name in parentheses. The locations Lawes, Brookstead, Nambour, and Redland Bay are all within 150 km of Brisbane, and cover a wide range of climatic and edaphic conditions, details of which are given in Shorter et al. (1977, p. 225). Before the trials started, it was anticipated that the performance of the lines would be somewhat similar at the two humid coastal locations, Nambour and Redland Bay, and that the performance at Lawes and Brookstead would be different from each other and from the two coastal locations. Redland Bay and Nambour were similar in that a soybean rust (*Phakopsara pachyrhizi*) epidemic occurred in both years of the test, although this was relatively more severe at Redland Bay in 1970, and less severe at that location in 1971. This disease occurred late in the season and had more effect on later-maturing lines. Lawes and Brookstead trials were free of this disease in both years of the test.

The experiment was a randomised complete block design with two replications in each location. A number of chemical and agronomic attributes were observed, but only the following are discussed here: seed yield (kg/ha), plant height (cm), lodging (rating scale 1–5), seed size (g/100 seeds), seed protein percentage, and seed oil percentage. Mungomery et al. (1974), Shorter et al. (1972), and

Table 1. Origin and maturity of soybean lines (after Mungomery et al., 1974)

Line no.	Name	Origin	Maturity[b]
1–40		Local selections[a]	9–11
43	CPI 17192 Mamloxi	Nigeria	11
41	CPI 15939 Avoyelles	Tanzania	9
42	CPI 15948 Hernon 49	Tanzania	9
45	Hampton	USA	8
48	Leslie	USA	8
49	Semstar	Local cultivar	8
50	Wills	USA	8
47	Jackson	USA	7
53	Bragg	USA	7
55	Lee	USA	6
56	Hood	USA	6
57	Ogden	USA	6
44	Dorman	USA	5
46	Hill	USA	5
54	Delmar	USA	4
58	Wayne	USA	3
51	CPI 26673	Morocco	3
52	CPI 26671	Morocco	3

[a] Local selections are derived from 41 (Avoyelles) and 43 (Mamloxi).

[b] Maturity is US maturity group classification or estimated equivalent.

Basford & McLachlan (1985) restrict their analyses to yield and protein percentage, while Basford (1982) discussed all six attributes.

Method of analysis

Traditionally genotypes have been characterised by an array of attributes producing a two-way table: the genotype × attribute (G × A) matrix. Alternatively, genotypes have been characterised by an array of performance values for a single attribute measured in a number of environments. This is a two-way table: the genotype × environment (G × E) matrix. The extension of these tables to the multi-attribute, multi-environment case produces a genotype × environment × attribute (G × E × A) matrix. As indicated earlier, the study of such three-way tables can potentially be of benefit to plant breeders, because they contain all the plant information from which inferences are to be made, as distinct from other measures on the environment.

Williams & Stephenson (1973) introduced a numerical method for the partition of three-dimensional data sets (sites × species × time) in marine ecology. Based on analysis of variance (equivalent to using Euclidean distance as a dissimilarity measure for classification), the 'mean variance per comparison' was used to assess the relative importance of dimensions or 'modes' and to provide a simple method of data reduction. Williams & Edye (1974) illustrated the applicability of this model to three-dimensional data matrices in agricultural experimentation, in particular they examined changes in botanical and chemical composition of pastures, i.e. their data were paddocks × measurements × time. Basford (1982) analysed the three-way genotype × environment × attribute matrix via individual differences scaling (see e.g. Carroll & Chang, 1970) by calculating for each environment the distances between genotypes from their (standardized) scores on the attributes. Effectively, this means that the (G × E × A) matrix with scores is transformed into a (G × G × E) matrix with distances. Another approach is that of Basford & McLachlan (1985) who considered a clustering of genotypes into groups based on the response in the other two modes, environments and attributes simultaneously. By appropriate specification of the underlying model, the mixture maximum likelihood method of clustering allows the (G × E × A) matrix to be handled directly.

In the present paper the (G × E × A) matrix will be analysed with three-mode principal component analysis (see e.g. Tucker, 1966; Kroonenberg & De Leeuw, 1980; Kroonenberg, 1983, 1984), which fits into the ordination rather than the clustering tradition. The aim of this procedure is to derive components for each of the ways or 'modes' (say, P, Q, R, of them for the first, second, third way or mode respectively), as well as a three-way matrix (the *core matrix*) of order P by Q by R. This core matrix G contains the weights assigned to each of the possible combinations of the components from the three modes. Thus g_{pqr} indicates the joint weight for the p-th component of the first mode, the q-th component of the second mode, and the r-th component of the third mode, and its squared value indicates the explained variation for that combination of components. The complete model may be written as

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip}\, b_{jq}\, c_{kr}\, g_{pqr} + e_{ijk}$$

with i=1, ..., 58 genotypes, j=1, ..., 8 environments, and k=1, ..., 6 attributes, and e_{ijk} the random error. An observed score x_{ijk} is thus 'modelled' as a systematic part of sums of multiplicative terms plus error. The a_{ip} are the entries of an I × P matrix A with the components for the first mode as its columns. The b_{jq} and c_{kr} are similarly defined for the second and third modes.

Supposing that clear-cut interpretations exist for the components in terms of latent entities, one way of interpreting the core matrix is to consider the elements g_{pqr} as the scores of (in our case) latent genotypes on latent attributes for latent environments (or types of environments). The g_{pqr} indicates the weight or importance of a particular combination of $a_{ip}b_{jq}c_{kr}$ for the modelling of x_{ijk}.

When such a clear-cut interpretation only exists for one mode, as is the case here for environments,

so-called joint plots can be made to investigate the relationships between each of the environment components and the original genotypes and attributes.

The program TUCKALS3 (Kroonenberg & Brouwer, 1985) was used to analyse the soybean data. This program is based on the alternating least squares algorithm described by Kroonenberg & De Leeuw (1980). Unlike the individual differences scaling reported by Basford (1982) this program handles only metric data.

For ease of interpretation, it is desirable to express the component configurations in low, preferably 2–4, dimensional space. However, representation of data in a reduced space inevitably results in some loss of information if the underlying spaces are of higher dimensionality. To assess the adequacy of the model the fitted sum of squares can be computed both for the overall solution and for each genotype, attribute, and environment separately (see Ten Berge et al., 1987). These fitted sums of squares can be expressed as squared multiple correlations between the data and their estimates based on the three-mode model.

Application

The data can be analysed in various ways depending on the focus or purpose of the research. One approach is to consider the data as a split-plot multivariate-multifactor design, in particular as six variates (attributes) with two independent variables, year (2 levels) and location (4 levels) as factors and the genotypes applied within each year-location combination (environment). The agronomist generally wants to investigate the main effects of overall quality and variability of locations over years, while plant breeders are especially interested in genotype by environment interactions for line selection purposes.

One of the major problems in using univariate and multivariate analysis of variance on such data is heterogeneity of error variances. Shorter (1972) and Mungomery (1978) investigated this aspect in depth for the current experiment. Analyses of variance for each attribute were computed and Tukey's test for additivity indicated that in general there was no reason to assume other than the usual additive model. Bartlett's test of homogeneity of variance across the eight environments indicated errors were heterogeneous. Various transformations were tested but resulted in little or no improvement in homogeneity except for the lodging score and seed size, and even there the test remained highly significant; the transformation, however, did not improve additivity in all environments. The major consequences of heterogeneity of error variances is on the test of the interaction mean square where too many significant tests are likely to occur. It seemed that this would be a serious problem only if the significance was marginal. The combined analyses over all environments were therefore computed using the untransformed data for all attributes, but taking into account that the error variances were heterogeneous when interpreting tests of significance (Shorter, 1972; Mungomery, 1978). A multivariate analysis of variance showed that the year main effect, the location main effect and the year by location interaction were significant. The same result applied for the univariate analyses, except for year and interaction effects for seed size. Thus the usual plant breeders' convention of identifying each year by location as an environment which influences plant response in a particular way was adopted.

Both multivariate and univariate F-tests with environments and genotypes as factors were significant. Table 2 gives the main effects for environment for each attribute. The main visual impression from this is that there is very little obvious pattern in the deviations or effects.

Variance among lines was partitioned into that attributable to within and between two groups. Group A consisted of the locally selected later maturing lines (1–43), while Group B was the largely introduced earlier maturing lines (44–58). Highly significant differences existed among lines within each group for all attributes except lodging score in the B group. The groups were significantly different for all but yield and lodging. Hence such a partition of variability was not very informative in explaining the pattern of plant response.

For each attribute k the data may be represented

via an additive linear model for the averages of the two replications per cell

$$x_{ij}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + \delta_{ij}^{(k)}$$

with i=1, ..., 58 genotypes; j=1, ..., 8 environments, k=1, ..., 6 attributes. Within plant-breeding research two common procedures for single attributes are employed – ordination and clustering (see Byth & Mungomery, 1981). Either $\alpha_i^{(k)} + \delta_{ij}^{(k)}$ or only the G × E interaction, $\delta_{ij}^{(k)}$ is used. In the present case, it was deemed important to relate differences in mean performance of genotypes to environment and attribute differences. Therefore, the first option was chosen, which means that $\hat{\mu}^{(k)} + \hat{\beta}_j^{(k)}$ are removed from the data.

The different units of measurement for the attributes make it imperative to equalise the scales per attribute before they can be analysed jointly, because otherwise there is no compatibility across attributes. Therefore, a scaling was performed over all genotype-environment combinations, so that the overall variability across attributes was equalised while maintaining the between-environment variability in the analysis. Because after scaling the interactions are comparable over attributes, in the sequel the index k will be written as any other index, i.e. as a subscript, rather than a superscript. More formally, if we define $\tilde{x}_{ijk}$ as

$$\tilde{x}_{ijk} = x_{ijk} - \hat{\mu}_k - \hat{\beta}_{jk}$$

where the carets indicate the usual least-squares estimators, then the scaling factors s are

$$s_k = [\sum_{i,j=1}^{I,J} (\tilde{x}_{ijk})^2]^{1/2}.$$

Model fit

Overall. Several solutions with different numbers of components for each of the modes were tried. Unfortunately, three-mode models are generally not nested, i.e. the size and nature of components may change when new components are added to the model. Therefore, several solutions have to be inspected to come to an adequate description of a data set. The squared multiple correlation for a solution with 3 components for genotypes, 2 for environments, and 2 for attributes, i.e. a 3×2×2-solution (Model I) was equal to 0.72. Alternatively, one may say that 72% of the variability measured by the uncorrected sum of squares of the data

Table 2. Main effects of environments for each attribute

Environments	Attributes[b]					
	Yield	Height	Lodging	Size	Protein	Oil
Lawes 1970	0.2	0.3	**1.3**	0.8	−0.8	0.8
Lawes 1971	0.5	−0.1	−0.3	0.2	−0.3	−0.3
Brookstead 1970	−0.5	0.1	0.0	−0.3	−0.2	−0.4
Brookstead 1971	0.4	0.1	0.4	1.4	1.4	−1.0
Nambour 1970	−0.2	−0.2	**−0.7**	0.7	**−3.6**	**2.8**
Nambour 1971	0.3	−0.3	**−1.0**	0.1	0.8	0.1
Redland Bay 1970	−0.4	0.0	0.6	−1.5	0.3	−0.9
Redland Bay 1971	−0.3	−0.0	−0.3	−1.5	**2.4**	−1.2
Attribute means	2.1	0.9	2.3	11.1	40.3	20.0
Standard error[a]	0.5	0.1	0.4	1.3	2.0	1.1

[a] Degrees of freedom for the standard errors is 399.

[b] The bold entries in the table are those effects which are different from all other effects for that attribute according to the Student-Newman-Keuls multiple range test.

could be fitted by the model. Adding a third component to attribute mode (thus fitting a 3×2×3 solution – Model II) increased the R^2 to 0.76. Subsequently, increasing the environment mode with a third component (3 × 3 × 3-solution – Model III) increased the R^2 to 0.77, while a 4 × 4 × 4-solution (Model IV) raises it to 0.81 at the cost of a large number of extra parameters and an increased complexity of interpretation. On the basis of informal judgements of the increases in R^2 compared to the increases in number of parameters and the interpretational qualities of the solutions, the 3 × 2 × 3-solution was deemed adequate and is reported here.

As a reviewer remaked, one would like to have more formal criteria for judging the adequacy of solutions. As far as we know, the only way to do this, is to assume the genotypes are random samples from some population (which they clearly are not, nor are they treated that way), because then the three-mode model can be reformulated as a regression model (see Kapteijn et al., 1986). For comparing two nested regression models under the assumption of independent and identically distributed errors with mean zero, an asymptomatic F-test is availablel, i.e.

$$(R_b^2 - R_a^2) / (1 - R_b^2)\} \star \{df_b - df_a)/(n - df_b)$$

where the subscript b refers to the less restricted and a to the more restricted model, and n is the number of observations (see e.g. Seber, 1977, p. 342). The F-statistics for the successive differences between the models are $F_{I,II}$ (7,2628) = 62.6, $F_{II,III}$ (12,2616) = 9.5, and $F_{III,IV}$(88,2528) = 6.1. Even though, all differences are very significant (which is largely due to n = 2784), only the comparison between Models I and II gives a really large F value. These tests concur with the informal conclusions above. Note, however, that hypothesis testing in this context is a rather dubious exercise.

Levels of modes. For eight of the 58 genotypes the model accounted for less than 35% of the variability in their response compared to the overall fit of 76%. In particular, these were the lines 2 (29%), 3 (32%), 24 (24%), 26 (11%), 27 (13%), 38 (12%), 41 (Avoyelles; 28%), and 42 (Hernon; 34%). All these genotypes, except for Avoyelles and Hernon, had generally low total variability indicating that they largely achieved average scores on the attributes in all environments. The comparatively low fit of 41 (Avoyelles) is somewhat surprising, as it is one of the two varieties from which the lines 1–40 were derived. The largest total variabilities were found for the non-local selections 45–58.

Even though there are some differences in fit between the environments and between the attributes, these are sufficiently small not to warrant a discussion.

Components description

Treating the components of the three modes separately gives only a partial view of the structure of the variability in the data. For a full view, it is necessary to look at the components of all modes simultaneously. As mentioned above the components of the genotypes (Table 3) and those of the attributes (Table 4) do not have obvious interpretations, and the lower-dimensional representations primarily serve the purpose of data reduction. We will, therefore, defer the discussion of the genotypes and attributes until later.

Environments. The two environment components partition the fitted variability into 71.5% and 4.5%, respectively. The first component (Table 5) is almost equal for all environments with the largest loadings for Redland Bay 70 & 71, and the smallest ones for Nambour 70 & 71. Thus this component reflects the overall similarity of the environments. The second component reflects a real Nambour – Redland Bay contrast, be it that Redland Bay 70 is rather extreme and that Lawes 70 joins Nambour on the other side of the component.

As in Basford (1982), the expectations expressed by Shorter et al. (1977, p. 225) about the similarity between the two coastal locations Nambour and Redland Bay is not supported by these outcomes, rather the opposite is true. Due to generally similar loadings, the first axis will be used to investigate the interactions between genotypes and attributes for

all environments together. The second component will be used to explore the differences between the two coastal locations, Nambour and Redland Bay.

Associative patterns of components

In this section the relationships between the reduced spaces of the three models will be addressed in two different ways. The first is to look at so-called joint plots, which portray the interactions between the genotypes and attributes for each of the components of the environments, and the second way is to look at the component scores of attribute-environment combinations on genotype components to focus more on the relationships between the attributes and environments, rather than on the genotypes.

Joint plots. Joint plots (a variant of Gabriel's (1971) biplot – see also Kroonenberg, 1985, p. 86, 87), display the relationships between genotypes and attributes for each environment component, i.e. they show what environments have in common (first joint plot – Fig. 1) and in which way Nambour and Redland Bay differ (second joint plot). The interpretation of such plots proceeds as for Gabriel's biplot based on the principle that distance in

Table 3. Genotype components

Genotype	Component			Genotype	Component		
	1	2	3		1	2	3
1	0.07	0.15	0.14	30	0.14	−0.16	0.02
2	0.03	0.07	0.04	31	0.08	0.03	0.07
3	−0.02	0.12	0.09	32	0.11	0.04	−0.23
4	−0.05	0.17	0.03	33	0.03	0.18	0.03
5	−0.06	0.17	0.13	34	0.10	0.06	0.01
6	−0.04	0.15	0.15	35	0.07	0.01	−0.01
7	−0.06	0.20	0.10	36	0.14	−0.14	0.04
8	−0.01	0.08	0.27	37	0.13	−0.05	−0.21
9	−0.05	0.13	0.15	38	0.02	0.01	0.05
10	−0.05	0.10	0.06	39	0.11	0.11	−0.18
11	0.15	−0.04	0.06	40	0.08	−0.04	−0.08
12	0.17	−0.08	−0.02	41 Avoyelles	0.06	0.09	−0.10
13	0.16	−0.06	−0.02	42 Hernon 49	0.01	−0.07	−0.28
14	0.08	0.17	0.11	43 Mamloxi	0.10	−0.10	0.13
15	0.06	0.05	0.04	44 Dorman	−0.15	−0.09	0.08
16	0.09	0.04	0.11	45 Hampton	−0.22	0.10	−0.07
17	0.18	−0.20	−0.07	46 Hill	−0.20	−0.03	0.12
18	0.19	−0.17	−0.07	47 Jackson	−0.21	−0.05	−0.09
19	0.13	−0.09	−0.04	48 Leslie	−0.19	0.13	−0.16
20	0.16	−0.16	−0.09	49 Semstar	−0.19	0.23	−0.17
21	0.14	0.02	0.09	50 Wills	−0.19	0.05	−0.20
22	0.16	−0.16	−0.13	51 CPI 26673	−0.15	−0.35	0.18
23	0.14	−0.04	0.00	52 CPI 26671	−0.15	−0.27	0.32
24	0.02	0.05	−0.07	53 Bragg	−0.21	0.00	−0.17
25	0.01	0.16	0.07	54 Delmar	−0.24	−0.13	0.08
26	0.01	0.03	−0.09	55 Lee	−0.22	−0.19	−0.10
27	0.00	0.05	−0.09	56 Hood	−0.21	−0.11	−0.17
28	0.07	0.18	0.12	57 Ogden	−0.21	−0.11	−0.24
29	0.09	−0.16	0.02	58 Wayne	−0.17	−0.27	0.23
Percentage variation accounted for					63.2	8.6	4.4

the plot is expressed through the inner product of two vectors. Two vectors are highly related if they are close together and thus have a high inner product, as for instance lodging and height in Figure 1a; they are unrelated if they are at right angles as for instance protein percentage and seed size; they are inversely related if they have angles of 180 degrees, as yield and protein.

To evaluate the importance of an attribute, say, protein percentage, for each genotype, one has to compare the projections of each genotype on the vector protein percentage. Similarly, one may compare the projections of the attributes on a genotype vector. In general, it is only necessary to look at one type of projection, and because of that, generally only the levels of one of the modes, here attributes, are indicated by vectors. The levels of the other mode are indicated by points, even though they are actually vectors. Returning to the protein percentage vector, it can be observed that of the non-local lines Morocco's 51 (CPI 26673) and 52 (CPI 26671) and 58 (Wayne) have the highest protein percentage (coupled with a moderately above average oil percentage), while the local cultivar 49 (Semstar) has a far below average protein percentage (but one of the highest oil percentages). Similarly, within the local selections (1–40, 41, 42, and 43) the major differences are especially due to differences in protein percentages of their seeds and their yields (with the attributes being inversely related), rather than for instance height and seed size.

Table 4. Environment components

Environments	E1	E2
Nambour 1970	0.23	**0.44**
Nambour 1971	0.29	0.32
Lawes 1970	**0.37**	**0.46**
Lawes 1971	**0.36**	0.09
Brookstead 1970	**0.35**	0.06
Brookstead 1971	**0.38**	−0.12
Redland Bay 1970	**0.43**	**−0.62**
Redland Bay 1971	**0.38**	−0.28
Percentage variation accounted for	71.5	4.6

Table 5. Attribute components

Attributes	A1	A2	A3
Oil percentage	**0.48**	−0.14	0.04
Seed size	**0.47**	0.32	0.34
Yield	0.33	**−0.57**	**0.59**
Protein percentage	**−0.36**	**0.50**	**0.70**
Lodging	**−0.40**	−0.23	0.12
Height	**−0.39**	**−0.49**	0.18
Percentage variation accounted for	60.8	10.8	4.5

Figure 1b shows a further 'refinement' of the differences in lines; it presents the first against the third axis, rather than the first against the second as in Figure 1a. There clearly exist differences between the very early, early (mid-)late maturing non-local lines. This is caused by the relatively lower yielding crop with relatively lower protein for the earlier lines compared to the later ones. Within the local selections this same pattern seems to be more related to individual genotypes, than to specific groupings of genotypes.

For comparison, in Figure 1 the grouping of genotypes resulting from the outcome of three-way cluster analysis of Basford & McLachlan (1985) is shown. The distinctiveness of clusters I, II, and III (non-local lines) and the other genotypes is evident. The clusters in the local selections clearly occupy different positions in the three-dimensional space, but some of the boundaries seem rather arbitrary. It is, however, comforting to see that the two methods support each other. In particular, the locations of the attribute vectors can be used to outline the principal differences between the clusters. For instance, cluster VII is particularly characterised by strong, tall plants with the highest protein percentage of the local selections but with a rather low yield. On the other hand, cluster IV genotypes are better characterised by considerable yields with above average oil and rather below average protein percentages. All the above statements can be made numeric by giving the actual values of the inner products of the vectors mentioned.

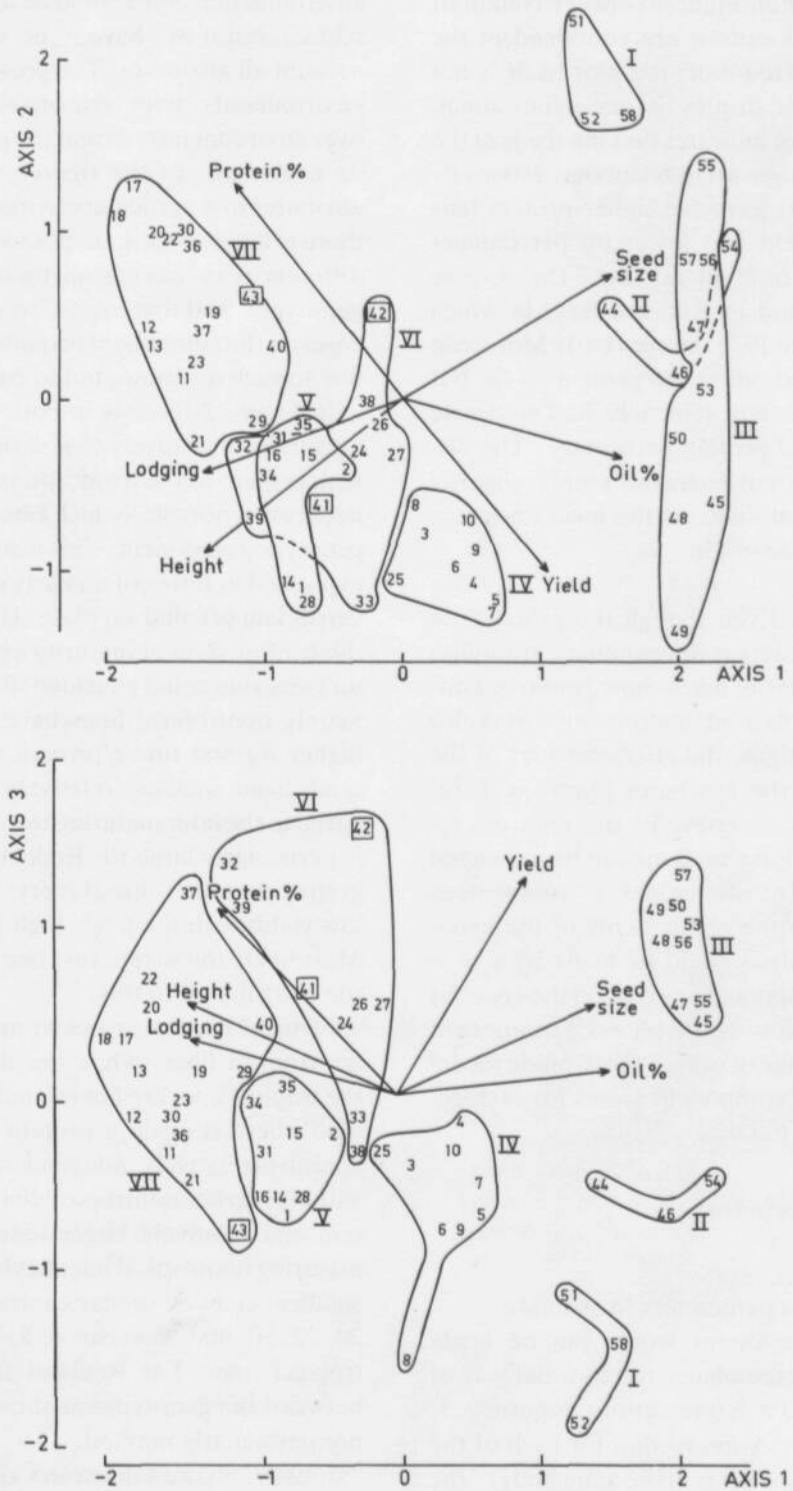

Fig. 1. Joint plot of genotypes and attributes for first environment component. (Arabic numerals refer to genotypes – see Table 1; Roman numerals refer to the clusters identified by Basford and McLachlan, 1985, Table 2; arrows indicate attributes).

Finally, the major differences between Nambour and Redland Bay locations are contained in the second joint plot. However, the plot itself is not shown, because the display is one-dimensional. The second joint plot indicates that the seeds of the non-local selections grown in Nambour, especially the very early ones, have far higher protein percentages, lower yield and lower oil percentages than those grown in Redland Bay. The reverse pattern can be found in Redland Bay, in which location especially in 1970 the very early Moroccan and Wayne lines had rather low protein levels, but high yields, and the local selections had moderate yields and increased protein percentage. The distinction between the environments seems primarily due to the non-local lines, as the local lines stay relatively close to the origin.

Component scores. Even though the primary interest of plant breeders is in examining groupings of genotypes in order to assess how genotypes differ in response to different environments, it is also important to investigate the characteristics of the environments and the attributes jointly with respect to genotype responses. To this end, the results of the three-mode analyses can be expanded to show the original attributes and environments as they are related to the components of the genotypes. Such an analysis could be made by a two-mode principal component analysis on the (A × E) by G matrix, and plotting the (A × E) component scores. The advantage of using a three-mode model to construct similar component scores for each genotype component p, i.e.

$$d_{pjk} = \sum_{q=1}^{Q} \sum_{r=1}^{R} b_{jq} c_{kr} g_{pqr},$$

is that there are less parameters to estimate.

Plots of the component scores can be made which bear some resemblance to the usual way of looking at plots of G × E interactions. Figures 2, 3, and 4 show the E × A interaction for each of the three genotype components in the main body of the figures, while along the right-hand vertical axis the component loadings of the genotypes for that component are schematically depicted. Possibly some environmental index could be used for the horizontal axis, but it will have to be an index taking into account all attributes. The present arrangement of environments gives reasonably smooth profiles over environments, so that the general patterns can be evaluated. In the figures large deviations of attributes in a particular environment indicate that there is large specific adaptation, and considerable differences in scoring on these attributes by the genotypes, and that the differences between genotypes on the component in question were especially due to such attributes in the environment.

In Figure 2 there is in both years and on most attributes a relatively low variability at Nambour with respect to the tropical-nontropical distinction between genotypes, which can be seen on the first genotype component. This distinction may also be expressed in terms of an early/mid (maturity: 3–6) versus late (9), and very late (11) difference, due to the confounding of maturity and origin. In the latter case, one could conclude that earlier maturing, mainly nontropical lines have higher yields with higher oil and lower protein percentages, larger seeds, and smaller, relatively weak plants compared to the later maturing tropical lines. The trend is particularly large for Redland Bay in 1970, suggesting that the tropical (very) late lines gave very low yields with relatively high protein percentage. Most likely the severe rust late in the growing season contributed to this.

Figure 3 illustrates a clear maturity effect in the nontropical lines, while the differences between the tropical lines are not related to maturity. In this case, there is again a protein percentage – yield contrast (relatively independent of oil percentage) with the earlier nontropical lines having more protein and relatively larger seeds, and the middle maturing nontropical lines having higher yields and smaller seeds. A similar contrast exists for 17, 18, 20, 22, 30, and 36 versus 4, 5, 7, 28, and 33 of the tropical lines. For Redland Bay the differences between the genotypes as shown in Figure 3 were not particularly marked.

Finally, Figure 4 illustrates again a contrast within both the nontropical and tropical lines, but in this case all environments produce relatively higher yields with higher protein percentages for the mid-

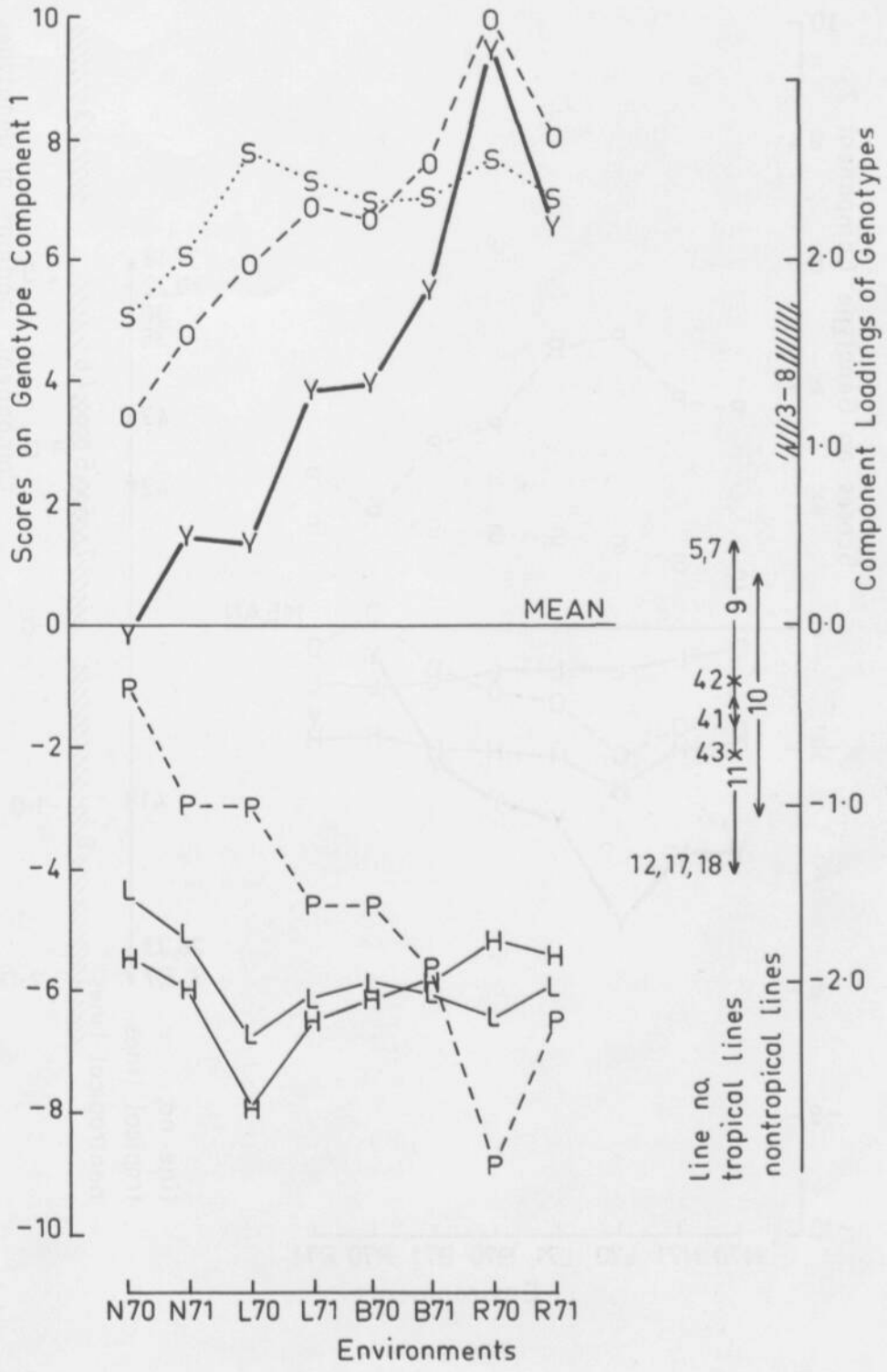

Fig. 2. Scores on genotype component 1 for all environment-attribute combinations. (S = seed size; O = oil percentage; Y = yield; N = Nambour; L = Lawes; B = Brookstead; R = Redland Bay; 70 = 1970; 71 = 1971. For line numbers see Table 1, for tropical and nontropical lines the approximate ranges of the maturity classes (3–11) are given for the genotype loadings on the right-hand side of the plot; ↔ approximate range tropical lines; //// or |||| approximate range nontropical lines. The zero MEAN is due to centring.)

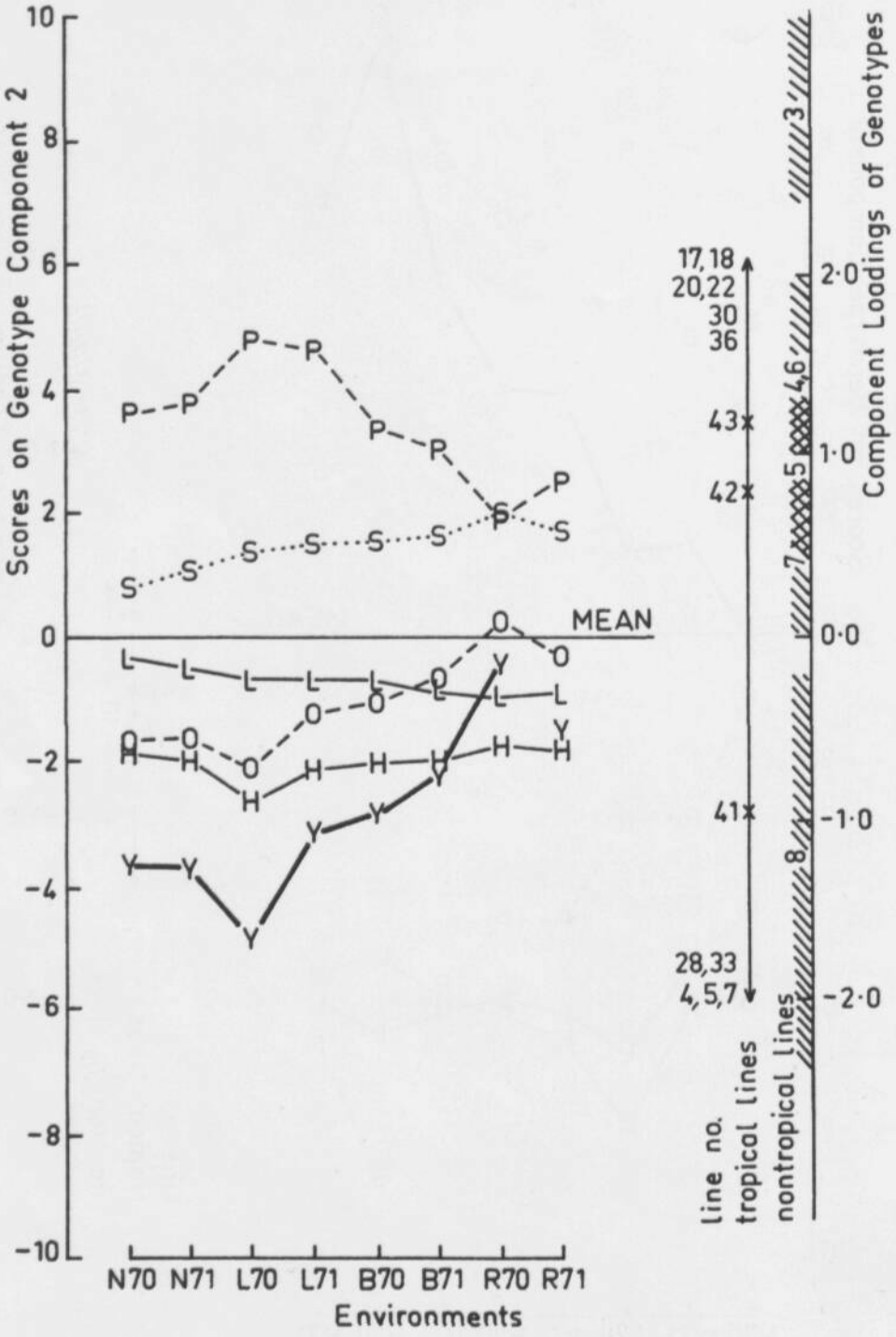

Fig. 3. Scores on genotype component 2 for all environment-attribute combinations. (For legend see below Fig. 2)

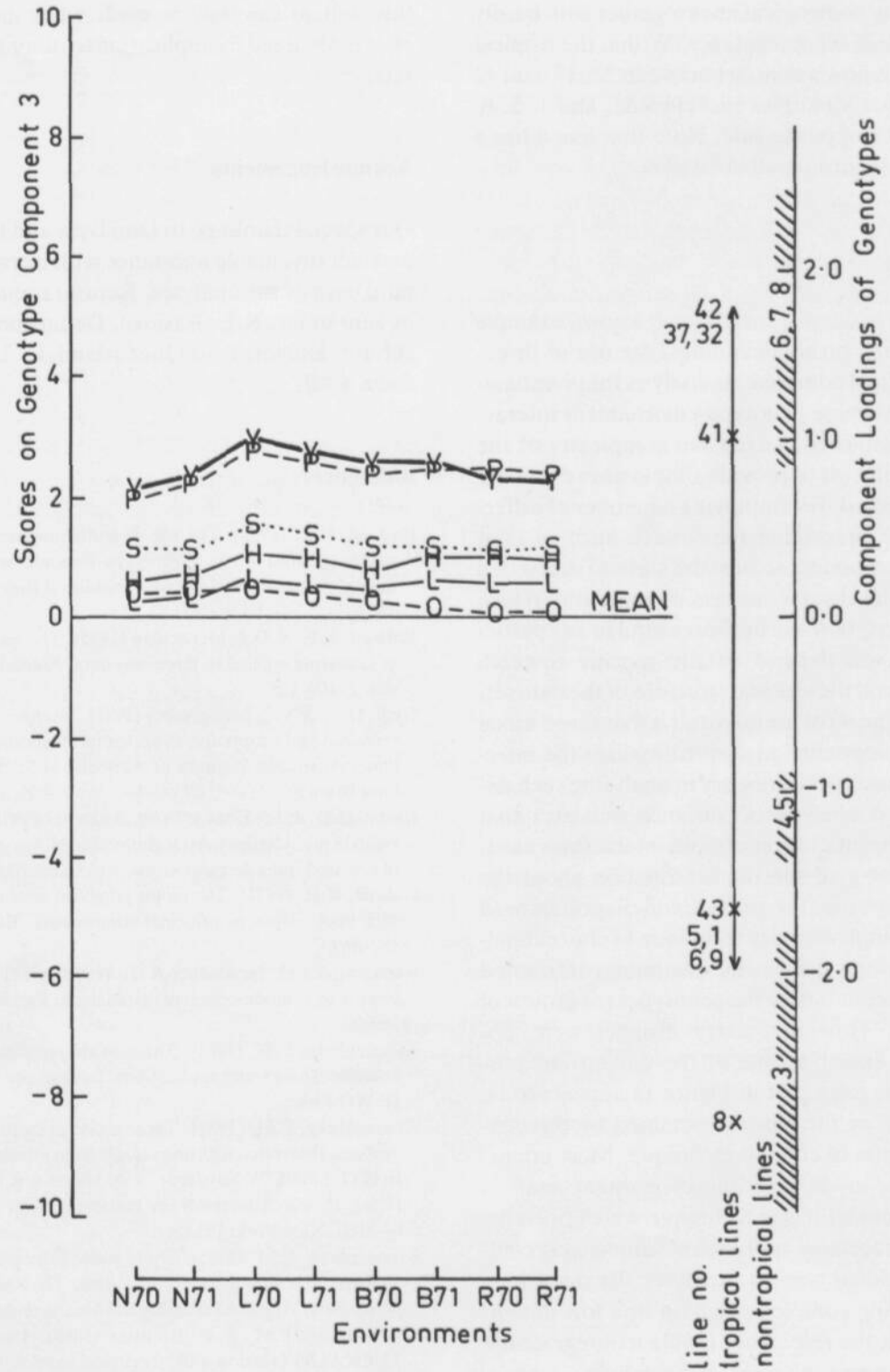

Fig. 4. Scores on genotype component 3 for all environment-attribute combinations. (For legend see below Fig. 2)

dle maturing nontropical lines together with hardly above average oil percentages. Within the tropical lines there is now a contrast between 32, 37, and 42 on the higher yield plus protein side, and 1, 5, 6, and 9 on the opposite side. Note that line 8 has a rather bad record on all attributes.

Conclusion

By treating in detail a specific well-known example from research on soybean lines, the use of three-mode principal components analysis for investigating multi-attribute genotype-environment interactions was explored. Given the complexity of the data, it is difficult to provide simple answers to the questions asked. By employing a number of different ways of presenting the results, such as joint plots and component scores, the method succeeded in illustrating diverse aspects of the data. Which type of description will be most useful in any particular study will depend on the specific research questions, and the size and structure of the data set.

Perhaps the most useful result is that three-mode principal component analysis formalizes the interpretative processes necessary in analysing such data. Standard analysis of variance indicates that many significant differences and interactions exist, but does not give specific information about the response patterns. The previous discussion showed it to be a complementary technique to cluster analysis in describing the way the attributes contributed to the differentiation of the genotypes (or groups of genotypes). However, extra insights were obtained, for example, one of the dimensions portrayed in the joint plot in Figure 1a appears to be independent of the clusters obtained by the mixture maximum likelihood technique. Most importantly, three-mode principle component analysis provides a model-based technique which prevents the rather piecemeal approach of subjectively combining individual two-way analyses. By describing the underlying complex situation in a low dimensional space, the researcher is able to integrate the response patterns inherent in the data in a reasonably direct manner. Any definitive recommendation of three-mode principal component analysis in this context can only be made when more experience is obtained by application to other similar data sets.

Acknowledgements

Our special thanks go to Don Byth and Ian DeLacy for their invaluable assistance with interpreting the outcomes of the analyses. Reprint requests should be sent to Dr. K.E. Basford, Department of Agriculture, University of Queensland, St. Lucia, Australia 4067.

References

Basford, K.E. (1982). The use of multidimensional scaling in analysing multi-attribute genotype response across environments. Australian Journal of Agricultural Research 33, 473–480.

Basford, K.E. & G.J. McLachlan (1985). The mixture method of clustering applied to three-way data. Journal of Classification 2, 109–125.

Byth, D.E. & V.E. Mungomery (1981). Interpretation of plant response and adaptation to agricultural environments. Brisbane: Australian Institute of Agricultural Science (Queensland Branch).

Carroll, J.D. & J.J. Chang (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35, 283–319.

Gabriel, K.R. (1971). The biplot graphical display of matrices with applications to principal components. Biometrika 58, 452–462.

Kapteyn, A., H. Neudecker & T. Wansbeek (1986). An approach to *n*-mode components analysis. Psychometrika 51, 269–275.

Kroonenberg, P.M. (1983). Three-mode principal component analysis: Theory and applications. Leiden, the Netherlands: DSWO Press.

Kroonenberg, P.M. (1984). Three-mode principal component analysis: Illustrated with an example from attachment theory. In H.G. Law, C.W. Snyder Jr., J.A. Hattie & R.P. McDonald (Eds), Research methods for multimode data analysis (pp. 64–103). New York: Praeger.

Kroonenberg, P.M. (1985). Three-mode principal component analysis of semantic differential data: The case of a triple personality. Applied Psychological Measurement 9, 83–94.

Kroonenberg, P.M. & P. Brouwer (1985). User's guide to TUCKALS3 (version 4.0). Technical report, University of Leiden, Department of Education.

Kroonenberg, P.M. & J. De Leeuw (1980). Principal component analysis for three-mode data by means of alternating

least squares algorithms. Psychometrika 45, 69–97.

Manning, H.L. (1956). Yield improvement from a selection index technique with cotton. Heredity 10, 303–322.

Mungomery, V.E. (1978). Genetic analyses of environmental interactions and effects of competition in soybeans. Unpublished Ph.D. thesis, Department of Agriculture, University of Queensland.

Mungomery, V.E., R. Shorter & D.E. Byth (1974). Genotype × environment interactions and environmental adaptation. I. Pattern analysis – application to soya bean populations. Australian Journal of Agricultural Research 25, 59–72.

Seber, G.A.F. (1977). Linear regression analysis. New York: Wiley.

Shorter, R. (1972). Influence of genotype and environment on chemical composition of soybean seeds (*Glycine max* (L.) Merrill). Unpublished M.Agr.Sc. thesis, University of Queensland.

Shorter, R., D.E. Byth & V.E. Mungomery (1977). Genotype × environment interactions and environmental adaptation. II. Assessment of environmental contributions. Australian Journal of Agricultural Research 28, 223–235.

Smith, H.F. (1936). A discriminant function for plant selection. Annals of Eugenics 7, 240–250.

Ten Berge, J.M.F., J. De Leeuw & P.M. Kroonenberg (1987). Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms. Psychometrika 52, 183–191.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. Psychometrika 31, 279–311.

Williams, W.T. & L.A. Edye (1974). A new method for the analysis of three-dimensional data matrices in agricultural experimentation. Australian Journal of Agricultural Research 25, 803–812.

Williams, W.T. & W. Stephenson (1973). The analysis of three-dimensional data (sites × species × times) in marine ecology. Journal of Experimental Marine Biology and Ecology 11, 207–227.