Journal of Classification 2:109-125 (1985)

Journal of
Classification

The Mixture Method of Clustering Applied to Three-Way Data

Kaye E. Basford
University of Queensland

Geoffrey J. McLachlan
University of Queensland

Abstract: Clustering or classifying individuals into groups such that there is relative homogeneity within the groups and heterogeneity between the groups is a problem which has been considered for many years. Most available clustering techniques are applicable only to a two-way data set, where one of the modes is to be partitioned into groups on the basis of the other mode. Suppose, however, that the data set is three-way. Then what is needed is a multivariate technique which will cluster one of the modes on the basis of both of the other modes simultaneously. It is shown that by appropriate specification of the underlying model, the mixture maximum likelihood approach to clustering can be applied in the context of a three-way table. It is illustrated using a soybean data set which consists of multiattribute measurements on a number of genotypes each grown in several environments. Although the problem is set in the framework of clustering genotypes, the technique is applicable to other types of three-way data sets.

Keywords: Clustering; Mixture maximum likelihood; Three-way data.

1. Introduction

The technique of clustering or classification uses the measurements on a set of elements to identify clusters or groups in which the elements are relatively homogeneous while they are heterogeneous between the clusters. The data itself may be in many forms and it is useful if a standard jargon for describing data sets can be established. In this paper we use the taxonomy of measurement data given by Carroll and Arabie (1980), where a mode is defined as a particular class of entities and an N-way array is defined as the cartesian product of a number of modes, some of which may be repeated. If the data are in the form of proximities between all the elements which we

Authors' Addresses: Kaye E. Basford and Geoffrey J. McLachlan, Department of Mathematics, University of Queensland, St. Lucia, 4067.

wish to cluster, then it would be described as one-mode two-way data. However, if the data consisted of the actual measurements of certain characteristics of the elements, then it would be described as two-mode two-way data. The latter is a more informative form of basic data set as it can easily be converted, if required, to the former by suitable definition of a similarity measure.

In most approaches to clustering in the biological sciences, the basic data is viewed as a two-mode two-way array, where one of the modes is to be partitioned into groups on the basis of the other mode. To illustrate this consider the data collected in a large plant improvement program, where an overall summarization of the patterns of genotype response is often more useful than the traditional comparison of individual responses. In such experiments two different types of two-mode two-way arrays are usually generated. The genotypes can be characterized by attributes producing a genotype by attribute $(G \times A)$ matrix. They can also be characterized by the performance values for a single attribute measured in a number of environments $(G \times E)$ matrix. Many methods of clustering have been applied to such two-way arrays to provide an appropriate grouping of the genotypes (Burt, Edye, Williams, Grof and Nicholson 1971; Mungomery, Shorter and Byth 1974; and Byth, Eisemann and De Lacy 1976). Such analyses have been very useful to plant breeders, but the restriction of being able to handle only two-way arrays has been a limitation.

Ideally, one would like to perform a clustering of the elements on the basis of all the information available assuming that differentiation between the groups is to be with respect to the total information. Thus in order to cluster the genotypes in the experiments defined above, it would be desirable to consider a combination of these two two-way arrays as a single three-way array. This produces a genotype by attribute by environment $(G \times A \times E)$ matrix which is a three-mode three-way data set. A clustering technique is required to group one of these modes (the genotypes) on the basis of both of the other modes (attributes and environments). Such an approach is beneficial for two reasons. Firstly, significant genotype by environment interaction is almost always present and it must be considered in the identification of groups of genotypes for which a general behavioral description is required. Secondly, it might be expected that in the underlying group structure the correlations between attributes would differ across groups of genotypes and a single attribute would not provide any information on this.

One way of developing a clustering technique appropriate to three-way data is to adapt the mixture maximum likelihood method of clustering, which is a model-based technique; see Hawkins, Muller and ten Krooden (1982, p. 353) for a discussion of this technique versus model-free approaches. The properties of the mixture method on data sets in the form of a two-way array have been studied by many authors, and a

comprehensive reference list has been given by McLachlan (1982). The applicability of the mixture method of clustering to three-way data is of considerable advantage as such arrays are produced in many disciplines.

The soybean data set chosen to illustrate this technique has been discussed in the literature before and the adaptation of the genotypes is well known (Mungomery et al. 1974; Shorter, Byth and Mungomery 1977; and Basford 1982). It permits some judgment on the usefulness of this method of clustering. Note that although the problem has been cast in the framework of multiattribute genotype responses across environments, this technique is applicable to other three-way data sets. One example would be an investigation aimed at grouping individuals based on the responses to different tests, say manual, intellectual and memory tests, taken under several blood alcohol levels.

2. Mixture Approach

Multivariate observations on a set of n elements forming a two-way array can be represented as $\mathbf{x}_1, \ldots, \mathbf{x}_n$. In applying the mixture method of clustering, it is assumed in the first instance that there is a specified number, say g, of underlying groups. A likelihood is then formed under the additional assumption that the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ has been drawn from a mixture of the underlying groups, designated here as $\Pi_1, \ldots, \Pi_g$, in some unknown proportions, $\pi_1, \ldots, \pi_g$, giving

$$L = \prod_{j=1}^{n} \{ \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j) \} \quad . \tag{1}$$

In the case of normality,

$$\mathbf{x}_j \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i) \quad in \quad \Pi_i (i = 1, \ldots, g) \quad , \tag{2}$$

and so $f_i(\mathbf{x}_j)$ in (1) refers to the normal density with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{V}_i$. The unknown parameters π_i, $\boldsymbol{\mu}_i$, and $\mathbf{V}_i (i = 1, \ldots, g)$ are estimated using the likelihood principle and, subsequently, each $\mathbf{x}_j$ can be allocated on the basis of the estimated posterior probabilities of group membership. The estimated posterior probability that element j with observation $\mathbf{x}_j$ belongs to Π_i, $\hat{\theta}_{ij}$, is formed by replacing the unknown parameters with their likelihood estimates in the expression for the true posterior probability, θ_{ij}. Thus

$$\hat{\theta}_{ij} = \hat{\pi}_i \hat{f}_i(\mathbf{x}_j) / \sum_{q=1}^{g} \hat{\pi}_q \hat{f}_q(\mathbf{x}_j), \quad (i = 1, \ldots, g) \ . \tag{3}$$

Element j is assigned to Π_q if

$$\hat{\theta}_{qj} > \hat{\theta}_{ij}, \quad i = 1, \ldots, g; \ i \neq q \ . \tag{4}$$

To apply this technique in the context of a three-way data set, the density function $f_i(\mathbf{x})$ of an element in population $\Pi_i (i = 1, \ldots, g)$ must include information on both of the other modes on which the measurements on the element were made. The particular example of clustering genotypes using genotype by attribute by environment data is considered here. The observation $\mathbf{x}_j (j = 1, \ldots, n)$ now contains the multiattribute responses of the jth genotype in all r environments, and is given by

$$\mathbf{x}_j = (\mathbf{x}'_{j1}, \ldots, \mathbf{x}'_{jr})' \ ,$$

where $\mathbf{x}_{jk}$ is a vector of length p giving the response of genotype j in environment k for each of the p measured attributes. With some applications there are replications and so $\mathbf{x}_j$ may represent an average over these replications. The underlying model corresponding to equation (2) is

$$\mathbf{x}_{jk} \sim N(\boldsymbol{\gamma}_{ik}, \mathbf{V}_i) \quad in \quad \Pi_i (i = 1, \ldots, g) \ , \tag{5}$$

where $\boldsymbol{\gamma}_{ik}$ is the mean response vector of group Π_i in environment k and $\mathbf{V}_i$ is the variance-covariance matrix of Π_i. This model covers the general situation where there may be some interaction between genotypes and environments; indeed, in the example to be presented there is a highly significant genotype by environment interaction. Under (5), the density of $\mathbf{x}_j$ in Π_i is equal to

$$f_i(\mathbf{x}_j) = (2\pi)^{-\frac{rp}{2}} |\mathbf{V}_i|^{-\frac{r}{2}} \exp\{-\frac{1}{2} \sum_{k=1}^{r} (\mathbf{x}_{jk} - \boldsymbol{\gamma}_{ik})' \mathbf{V}_i^{-1} (\mathbf{x}_{jk} - \boldsymbol{\gamma}_{ik})\} \ . \tag{6}$$

The likelihood estimates of the unknown parameters can be expressed subsequently as

$$\hat{\pi}_i = \sum_j \hat{\theta}_{ij}/n \ , \quad (i = 1, \ldots, g) \quad , \tag{7}$$

$$\hat{\gamma}_{ik} = \sum_j \hat{\theta}_{ij} \mathbf{x}_{jk} / \ (n\hat{\pi}_i) \ , \quad (i = 1, \ldots, g) \quad , \tag{8}$$

$$\hat{\mathbf{V}}_i = \sum_{jk} \hat{\theta}_{ij} (\mathbf{x}_{jk} - \hat{\gamma}_{ik}) \ (\mathbf{x}_{jk} - \hat{\gamma}_{ik})' / \ (nr\hat{\pi}_i) \ , \quad (i = 1, \ldots, g) \quad . \tag{9}$$

The estimated posterior probabilities $\hat{\theta}_{ij}$ have the same form as in equation (3), but where now $f_i(\mathbf{x}_j)$ is given by equation (6).

The computation of the likelihood estimates is facilitated by identifying these equations with the application of the EM algorithm of Dempster, Laird, and Rubin (1977). For given starting values of the parameters the expectation and maximization steps of this algorithm are alternated until convergence in the case of a sequence of likelihood values bounded above. The process should be repeated for various starting values in an attempt to locate all local maxima of the likelihood.

The estimate of the ith group mean over all environments is given by

$$\begin{aligned} \hat{\boldsymbol{\mu}}_i &= \sum_k \hat{\gamma}_{ik} / \ r \\ &= \sum_j \hat{\theta}_{ij} \ \bar{\mathbf{x}}_j \ . / \ (n\hat{\pi}_i), \quad (i = 1 \ , \ldots \ , \ g) \quad . \end{aligned} \tag{10}$$

Thus we can write

$$\hat{\gamma}_{ik} = \hat{\boldsymbol{\mu}}_i + \sum_j \hat{\theta}_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j \ .) / \ (n\hat{\pi}_i) \quad , \tag{11}$$

where the second term on the right-hand side of (11) represents the deviation from the expected mean response for Π_i. This deviation can be considered as the sum of the effect of the kth environment and the interaction between the ith group and the kth environment.

3. Number of Clusters

Testing for the existence of different clusters is an important but difficult problem. Unfortunately, the likelihood ratio criterion for testing the hypothesis of g_1 versus g_2 populations $(g_1 < g_2)$ does not have its usual

asymptotic distribution (Hartigan 1977 and Binder 1978). We follow here the suggestion of Wolfe (1971) that under the null hypothesis

$$-2C \log \lambda \sim \chi^2_d \tag{12}$$

approximately, where d, the degrees of freedom of the chi-square distribution, is taken to be twice the difference in the number of parameters in the two hypotheses not including the proportions. A comprehensive account of the accuracy of this approximation may be found in Basford and McLachlan (1985). The constant C is introduced to improve the approximation. In the present context we take

$$C = \{r\ (n-1) - \frac{1}{2}\ (p + g_2)\}/\ nr \tag{13}$$

corresponding to Kendall (1965, p. 134) who applied Bartlett's approximation to the problem of testing between g_2 means in a multivariate analysis of variance by first eliminating the block differences from the variation. The r blocks in Kendall's design can be considered analogous to the r environments here.

The reliability of the approximation (12) will, of course, depend on the size of the sample. Therefore the outcome of the likelihood ratio test should not be rigidly interpreted, but rather used as a guide to the possible number of underlying groups. Examination of the posterior probabilities of group membership for the genotypes for values of g near to the value accepted according to the likelihood ratio test can be useful in leading to the final decision on the number of groups.

4. Experimental Details

Mungomery et al. (1974) first reported the experiment from which this data set was 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 locations, Redland Bay, Lawes, Brookstead and Nambour, were within 150 km of Brisbane, and covered a wide range of climatic and edaphic conditions. The experiment was a randomized complete block design with two replicates in each environment; further details of the experimental procedures are given by Mungomery et al. (1974). A number of chemical and agronomic attributes was observed, including seed yield (kg/ha), plant height (cm), lodging (rating scale 1-5), seed size (g/100 seeds), seed protein percentage and seed oil percentage. We shall focus attention here on two of these, seed yield and seed protein

TABLE 1

Origin and Maturity of Soybean Lines tested across Four Locations in Each of 2 Years (after Mungomery et al. 1974)

Line No.	Name	Origin	Relative maturity
1-40		LS[A]	Mid-very late (8-11)[B]
41	CPI 15939 Avoyelles	Tanzania	Late-mid (9)
42	CPI 15948 Hernon 49	Tanzania	Late-mid (9)
43	CPI 17192 Mamloxi	Nigeria	Very late (11)
44	Dorman	U.S.A.	Early (5)
45	Hampton	U.S.A.	Mid (8)
46	Hill	U.S.A.	Early (5)
47	Jackson	U.S.A.	Early-mid (7)
48	Leslie	U.S.A.	Mid (8)
49	Semstar	Local cultivar	Mid-late (8)
50	Wills	U.S.A.	Mid (8)
51	CPI 26673	Morocco	Very early (3)
52	CPI 26671	Morocco	Very early (3)
53	Bragg	U.S.A.	Mid (7)
54	Delmar	U.S.A.	Early (4)
55	Lee	U.S.A.	Early-mid (6)
56	Hood	U.S.A.	Early-mid (6)
57	Ogden	U.S.A.	Early-mid (6)
58	Wayne	U.S.A.	Very early (3)

[A]LS, local selections from Mamloxi (CPI 17192) x Avoyelles (CPI 15939).

[B]Number in parentheses is US. maturity group classification or estimated equivalent.

percentage. A clustering of the soybean lines using each of these two attributes separately has been reported by Mungomery et al. (1974). The mixture approach can be applied for $p>2$ attributes, but as p increases, the number of parameters in the model (5) increases sharply, greatly compounding the problems with multiple maxima. A clustering of the genotypes using the responses of all six attributes at each location in each year was undertaken according to the mixture method. The same number of underlying groups was suggested with only slightly different group composition to that obtained using just seed yield and seed protein percentage, as detailed below. Because the pattern of group response is perhaps best interpreted by graphs of expected response for each attribute it was decided to illustrate this method of clustering for $p = 2$ attributes. As each attribute was observed in four sites in two successive years, there were eight effective environments. There were two replications in each environment and, as with the analyses of Mungomery et al. (1974), the basic data set $\mathbf{x}_{jk}$ is taken to be the mean response over the replicates in each environment.

5. Application of Mixture Approach

The mixture maximum likelihood method of clustering was applied to the soybean data set to obtain groups of genotypes within which there were similar behavioral response patterns. As recommended earlier, several starting values were used for each value of g, the specified number of underlying groups. An initial grouping of the genotypes can be obtained by focusing attention on a single attribute and using the corresponding analysis of variance table and subsequent multiple comparisons of genotype means. Alternatively, initial groupings can be obtained by using the results of other clustering techniques applied to the genotype by environment data for a single attribute. Both these methods were tried here.

The likelihood increased substantially with increasing g but flattened out after $g = 7$. For testing $g = 6$ against $g = 7$ the P value according to the approximation (12) was less than 0.05 while for testing $g = 7$ against $g = 8$ the P value was 0.32. Investigation of posterior probabilities for values of g close to seven provided further support for $g = 7$. In accordance with (4), each genotype was allocated to the group to which it had the highest estimated posterior probability of belonging. As the smallest maximum value was 0.916, it would appear that the genotypes can be clustered with a high degree of certainty. The identification of the soybean lines forming the groups is given in Table 2. The estimated means, calculated from equation (10), for seed yield and protein percentage, in addition to the correlation coefficient between these attributes, as determined from the estimated covariance matrices, are reported in Table 3.

The mixture approach has been implemented here with an arbitrary covariance matrix $\mathbf{V}_i$ within each group i. However, Aitkin, Anderson and

TABLE 2

Identification of the Soybean Lines forming the Groups obtained by the Mixture Method of Clustering

Group	Line numbers
I	51, 52, 58
II	44, 46, 54
III	45, 47, 48, 49, 50, 53, 55, 56, 57
IV	3, 4, 5, 6, 7, 8, 9, 10, 25
V	1, 2, 14, 15, 16, 28, 31, 34, 35
VI	24, 26, 27, 32, 33, 38, 39, 41, 42
VII	11, 12, 13, 17, 18, 19, 20, 21, 22, 23, 29, 30, 36, 37, 40, 43

TABLE 3

Estimated Mean Effect for each Attribute and Correlation Coefficient within the Groups

Group	Yield (kg/ha)	Protein Percentage	Correlation Coefficient
I	1451.4	39.5	-0.47
II	2227.0	38.1	0.07
III	2879.2	38.9	-0.24
IV	2206.2	38.1	0.05
V	1899.1	40.1	-0.13
VI	2191.7	41.0	-0.04
VII	1566.3	42.7	-0.08

Hinde (1981) suggested that in many instances it may be reasonable to assume conditional independence; that is, the observed correlations between the attributes result from the clustered nature of the sample, and that within

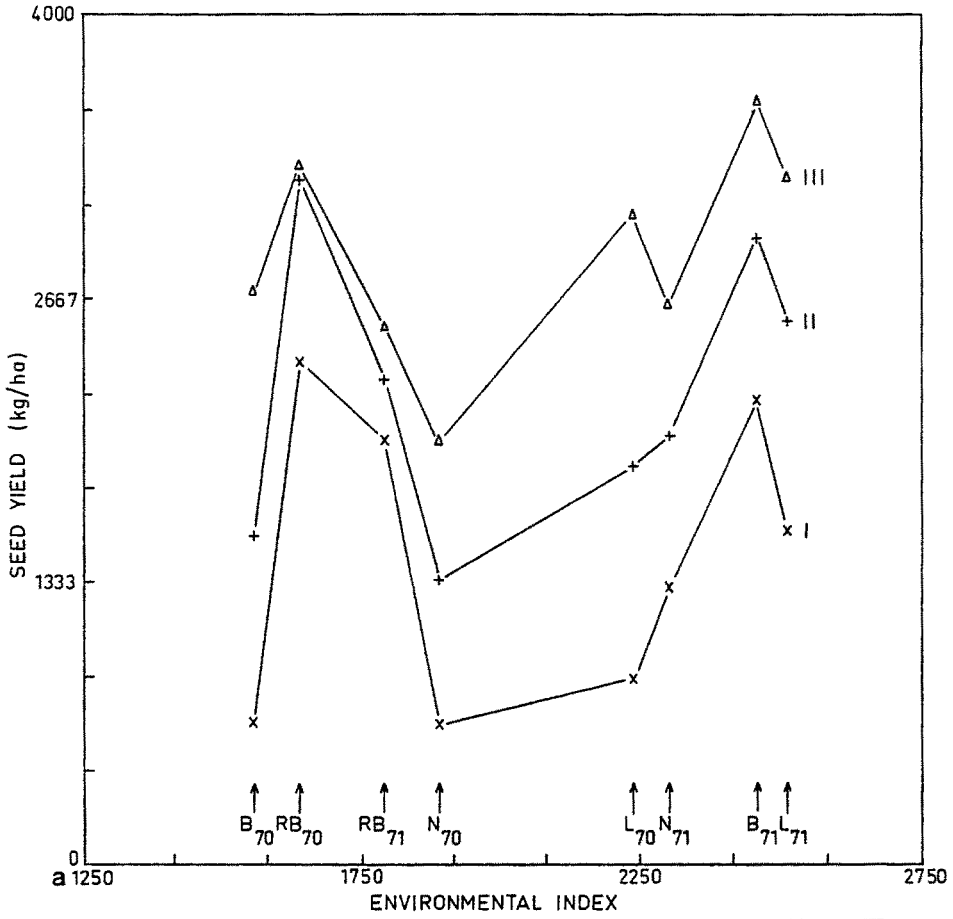

Figure 1a. Plot of estimated expected seed yield in each of groups I, II and III against each environment mean over all genotypes.

the underlying groups there is zero correlation between the attributes. The present example appears to show some support for this proposition for the estimated correlation coefficient between yield and protein percentage within a group (see Table 3) is generally quite small. Only in one case, group I, is it larger, in absolute terms, than the overall sample value of −0.34. An analysis of the data under the restricted model of a diagonal covariance matrix within each group (that is, under the assumption of independence of the attributes within a group), was undertaken and resulted in a somewhat different clustering of the genotypes. A five group rather than a seven group description of the data set was concluded to be adequate, although group composition showed considerable resemblance to part of that displayed in Table 2.

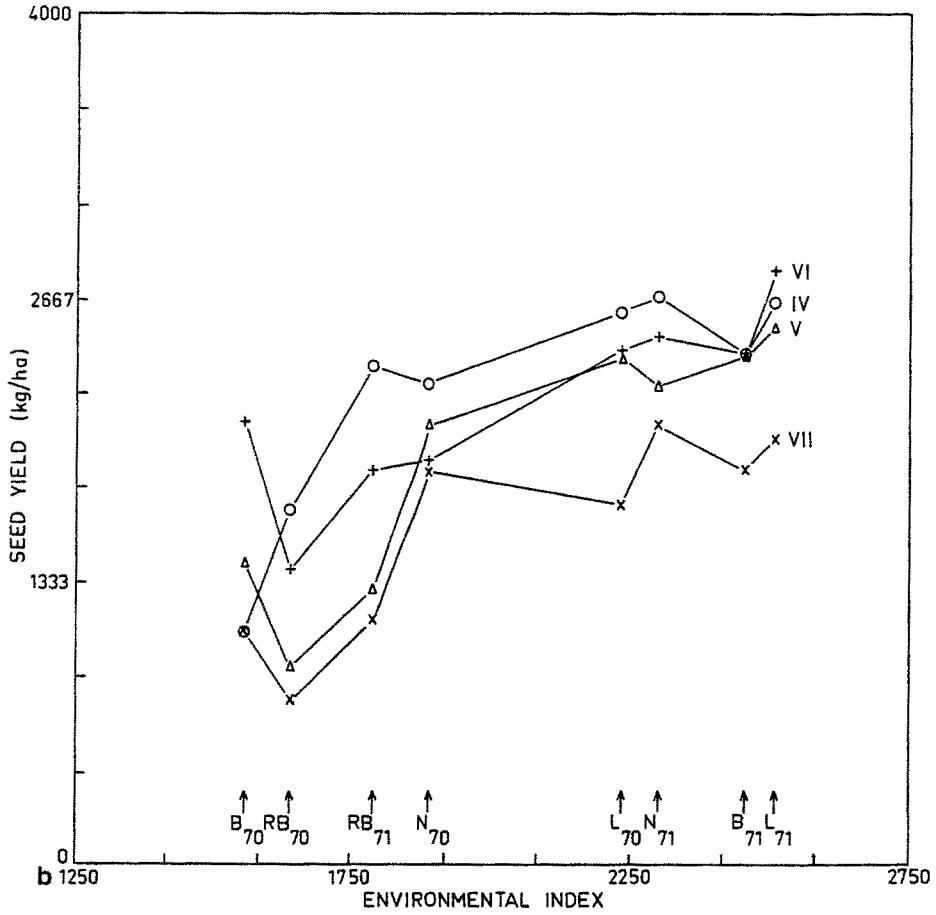

Figure 1b. Plot of estimated seed yield in each of groups IV, V, VI and VII against each environment mean over all genotypes.

6. Interpretation of Results

In Table 3, we display the values of $\hat{\mu}_i$, the estimated mean vector of the attributes over all environments in the ith group. However, to explain the differences between the groups, we need to consider the pattern of group responses across environments as exhibited in Figures 1 and 2. In these figures, the estimated expected response, $\hat{\gamma}_{ik}$, for the ith genotype group in the kth environment is plotted against the kth environment mean over all genotypes for seed yield and protein percentage respectively. In each case, the horizontal axis is an index of increasing environmental response for that attribute. The environments are denoted by the site initials and year subscript; for example, RB_{70} refers to Redland Bay in 1970.

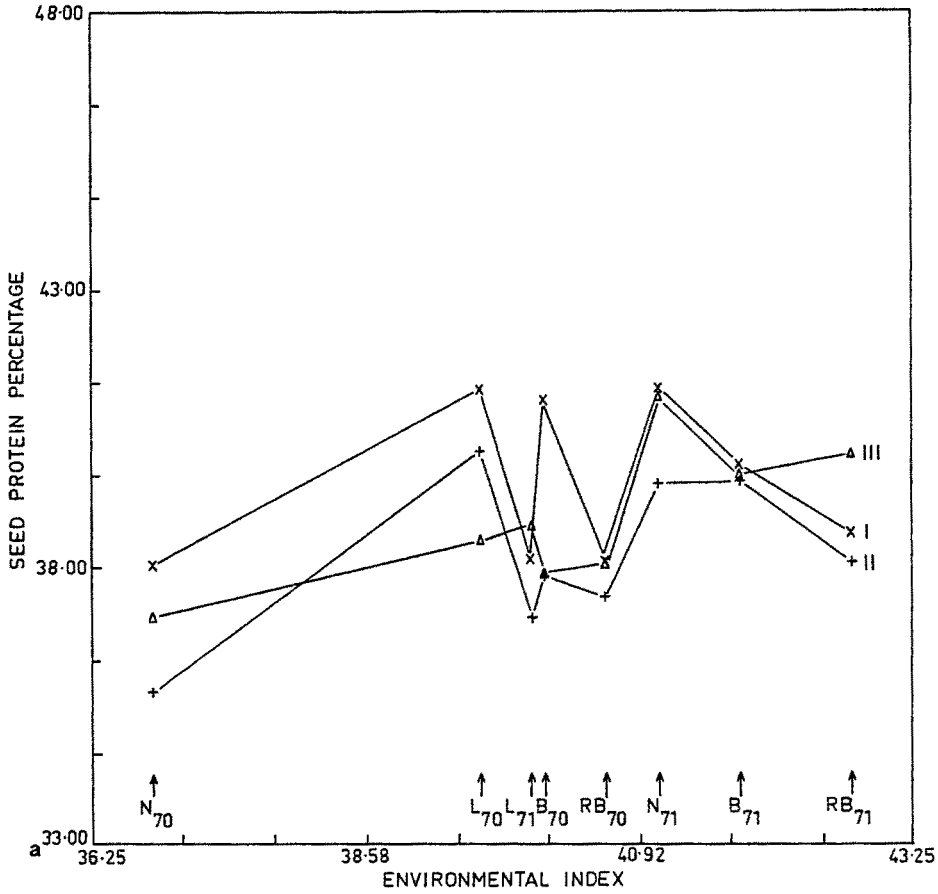

Figure 2a. Plot of estimated expected protein percentage in each of groups I, II and III against environment mean over all genotypes.

From inspection of Figures 1 and 2, it is apparent that on the basis of response pattern these groups can be viewed as two subsets: (a) containing groups labeled I to III and (b) containing groups labeled IV to VII. The patterns are particularly alike within subset (a) for yield and within subset (b) for protein percentage. Subset (a) contains those lines with early to mid maturity group classification or estimated equivalent while subset (b) contains lines 41, 43, all progeny from their cross and only one other line, 42, which is also late maturing (Table 1).

It can be seen from Figure 1 that at any given environment seed yield increases through groups I, II and III. For each group in this subset, the highest yield was recorded at either B_{71} or RB_{70} where group II had a similar yield to group III. In subset (b), group VII had the lowest seed yield at

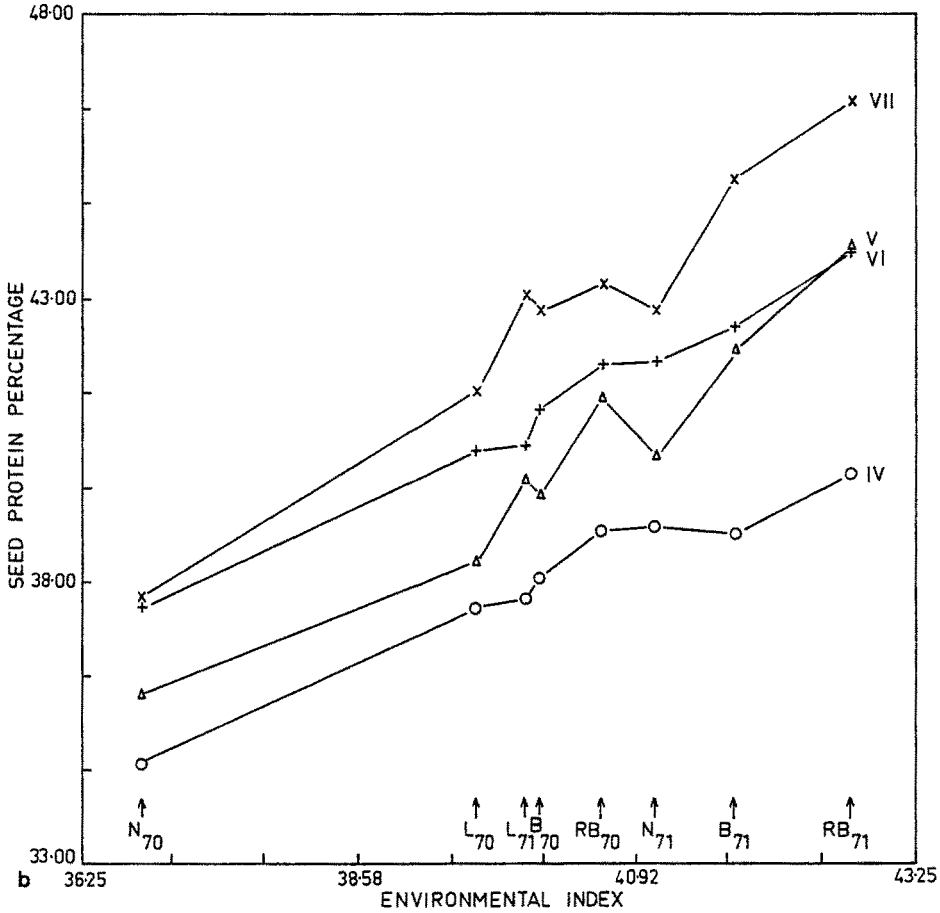

Figure 2b. Plot of estimated expected protein percentage in each of groups IV, V, VI and VII against environment mean over all genotypes.

all except one environment, while group IV had the highest seed yield at most of the environments. Seed yield for subset (b) was higher at L_{71} and N_{71} than at the other environments. The environment which best distinguished between subsets (a) and (b) in terms of seed yield was RB_{70} at which the former subset had much greater yields. At RB_{71}, however, the responses were much more similar. By contrast, seed yields for the other three sites increased from 1970 to 1971. It can be seen from Figure 2 that within subset (a) the more favorable environments for protein percentage were L_{70} and N_{71}; within subset (b) the most favorable environment was RB_{71}. This environment, RB_{71}, also provided the widest range in protein percentage over the groups while there was far less spread at N_{70} or N_{71}. Generally, protein percentage increased through groups IV, V, VI and VII at all environments, starting with the lowest values at N_{70}. Protein percentage

was greater in 1971 than in 1970 in all groups at all sites except at Brookstead and Lawes where groups I and I and II, respectively, had lower protein percentage.

7. Discussion

The brief summarization of the response patterns as modeled under the mixture approach illustrates how useful this method of clustering can be. By incorporating both attributes and environments, an overall picture of response is used to group the genotypes. To understand and explain the group response it was beneficial here to draw a graph of expected response of each group in each environment for each attribute. Mungomery et al. (1974) produced such graphs of mean group response in each environment but as independent clusterings were obtained, different groupings were plotted for each attribute. By simultaneously using both attributes in the clustering technique, the problem does not arise of trying to reconcile the different groupings which may be obtained using each attribute separately. In the present analysis, for instance, not only did a clearly interpretable separation based on mean effect of the early maturing lines, predominantly of United States origin, in groups I to III for yield occur, but also such a separation of the locally selected later maturing lines for protein percentage was observed in groups IV to VII.

It has been noted in the Introduction that plant breeders are interested in homogeneous groups of genotypes, particularly for a convenient summary of response patterns. There has been some discussion in the literature on combining attributes to produce a biologically meaningful measure. For instance, seed yield and seed protein percentage could be combined to form a new attribute called seed protein yield. Selection indices are another way of combining attributes into a single measure. In many cases, however, no obvious or appropriate measure is available. Thus if the clustering is to be based on the information inherent in all the attributes measured in each environment then the proposed technique must be able to handle three-mode three-way data. It was decided not to consider any of the variable reduction techniques such as principal component analysis or generalized canonical correlation to convert the data set to a two-way array and thereby permit analysis with a conventional clustering technique. These appeared to be circumventing the problem of determining a method of clustering to analyze data directly in the form of a three-mode three-way array. Also, Chang (1983) showed that the practice of applying principal components to reduce the dimension of the data before clustering is not justified in general. By considering a mixture of two multivariate normal distributions with a common covariance matrix, he showed that the components with the larger eigenvalues do not necessarily contain more information on the distance between groups.

Carroll and Arabie (1983) devised a method for non-hierarchical overlapping clustering called INDCLUS for the case of three-way proximity data. Data in the form of a three-mode three-way array are first converted, using a similarity measure, to two-mode three-way data. This produces a matrix of element by element proximity measures for each of a number of individual subjects or data sources. Carroll, Clark and De Sarbo (1984) developed a new methodology called INDTREES for fitting a hierarchical tree structure to obtain a discrete network representation of such proximity data. The individual differences generalization is one in which individuals, for example, are assumed to base their judgments on the same family of trees, but are allowed to have different node heights and/or branch lengths. Their method minimizes the sum of the squared differences between fitted and observed dissimilarity between elements within a data source, while satisfying the ultrametric inequality. That is, given two disjoint clusters, all distances between elements in the same cluster are smaller than distances between elements in two different clusters, and that these between-cluster distances are equal (Carroll et al. 1984). It would be possible to apply such methods to the present problem by calculating a dissimilarity measure between genotypes within each environment and considering each environment as an individual data source. Squared Euclidean distance could be used as the proximity measure, but other choices are available.

Recently, De Sarbo, Carroll, Clark and Green (1984) proposed a new clustering method, called SYNCLUS, for clustering elements on which a battery of variables has been measured. It is an algorithm for K-means clustering (MacQueen 1967) using a weighted mean-square, stress-like measure, and can be generalized to handle three-way data. SYNCLUS can be applied in those situations where it is appropriate to put prior weighting on particular batteries of variables and then allow the clustering procedure to weight the variables within these batteries according to their relative importance to the clustering (De Sarbo et al. 1984). With respect to the present problem, the elements would be the genotypes and the attributes measured in each environment would be the variables in each battery. It is possible that the genotype by environment interaction might be expressed in the SYNCLUS model by different weightings on the attributes in each environment, but it might not be straightforward to interpret these weightings in terms of the interaction. Thus it was felt more appropriate to use a method of clustering which incorporated this genotype by environment interaction directly into the underlying model.

In analyzing such three-way data, Basford (1982) considered a multidimensional scaling (MDS) approach to obtain a spatial representation in a low dimensional space (Shepard 1962a, 1962b; Kruskal 1964a, 1964b). The relative proximity of the points (genotypes in this instance) in this space was then used as an indication of similarity of response pattern. Kruskal (1977), Whitmore and Harner (1980), and Morgan (1981) have all noted that, in

general, a good overall picture is obtained using MDS but that it is not as sensitive to local features of the arrangement. Ramsey (1982) reviewed the statistical problems associated with MDS and in the subsequent discussion the exploratory nature of this graphical technique was stressed. It should be clear therefore that MDS is not a competing technique but rather a complementary one to clustering (Kruskal 1977).

In addition to the empirical example analyzed, a Monte Carlo experiment was conducted to assess the performance of the mixture method in identifying a known group structure. A simulated data set was obtained by generating observations $\mathbf{x}_{jk}(j = 1, \ldots, 50; k = 1, \ldots, 8)$ from a mixture in equal proportions of $g = 4$ bivariate normal distributions in accordance with model (5). The parameter values $\boldsymbol{\gamma}_{ik}$ and $\mathbf{V}_i$ were set equal to the estimates of the corresponding parameters obtained for the four groups labeled IV to VII in the seven group solution of the real example discussed above. An IMSL subroutine based on the inverse method was used to generate normal random variables from uniformly distributed densities which were produced by a multiplicative congruential generator of the form $X_{i+1} \equiv r\ X_i (modulos)$, where $r = 7^5$ and $s = 2^{31}-1$. In all, 25 such sets were generated in this manner and for each set the mixture method of clustering was applied and its correct allocation rates noted. The overall correct allocation rate averaged over the 25 simulations was equal to 0.76. This is an encouraging result, considering that there was a considerable degree of overlap between the underlying simulated groups as evidenced from the generated data.

References

AITKIN, M., ANDERSON, D., and HINDE, J. (1981), "Statistical Modelling of Data on Teaching Styles," *Journal of the Royal Statistical Society, A 144*, 419-461.

BASFORD, K.E. (1982), "The Use of Multidimensional Scaling in Analysing Multi-attribute Genotype Response Across Environments," *Australian Journal of Agricultural Rsearch, 33*, 473-480.

BASFORD, K.E., and MCLACHLAN, G.J. (1985), "Cluster Analysis in a Randomized Complete Block Design," To appear in *Communications in Statistics - Theory and Methods.*

BINDER, D.A. (1978), "Bayesian Cluster Analysis," *Biometrika, 65*, 31-38.

BURT, R.L., EDYE, L.A., WILLIAMS, W.T., GROF, B., and NICHOLSON, C.H.L. (1971), "Numerical Analysis of Variation Patterns in the Genus Stylosanthes as an Aid to Plant Introduction and Assessment," *Australian Journal of Agricultural Research, 22*, 737-757.

BYTH, D.E., EISEMANN, R.L., and DE LACY, I.H. (1976), "Two-way Pattern Analysis of a Large Data Set to Evaluate Genotype Adaptation," *Heredity, 37*, 215-230.

CARROLL, J.D., and ARABIE, P. (1980), "Multidimensional Scaling," *Annual Review of Psychology, 31*, 607-649.

CARROLL, J.D., and ARABIE, P. (1983), "INDCLUS: An Individual Differences Generalization of the ADCLUS Model and the MAPCLUS Algorithm," *Psychometrika, 48*, 157-169.

CARROLL, J.D., CLARK, L.A., and DE SARBO, W.S. (1984), "The Representation of Three-way Proximity Data by Single and Multiple Tree Structure Models," *Journal of Classification, 1*, 25-74.

CHANG, W.C. (1983), "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions," *Applied Statistics, 32*, 267-275.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, B 39*, 1-38.

DE SARBO, W.S., CARROLL, J.D., CLARK, L.A., and GREEN, P.E. (1984), "Synthesized Clustering: A Method for Amalgamating Alternative Clustering Bases with Differential Weighting of Variables," *Psychometrika, 49*, 57-78.

HARTIGAN, J.A. (1977), "Distribution Problems in Clustering," In *Classifications and Clustering*, (Ed.) J. van Ryzin, New York: Academic Press, 45-71.

HAWKINS, D.M., MULLER, M.W., and TEN KROODEN, J.A. (1982), "Cluster Analysis," In *Topics in Applied Multivariate Analysis*, (Ed.) D. M. Hawkins, Cambridge: Cambridge University Press, 303-356.

KENDALL, M.G. (1965), *A Course in Multivariate Analysis*, London: Charles Griffin.

KRUSKAL, J.B. (1964a), "Mulidimensional Scaling by Optimizing Gooness-of-Fit to a Nonmetric Hypothesis," *Psychometrika, 29*, 1-27.

KRUSKAL, J.B. (1964b), "Nonmetric Multidimensional Scaling," *Psychometrika, 29*, 115-129.

KRUSKAL, J.B. (1977), "The Relationship Between Multidimensional Scaling and Clustering," In *Classification and Clustering, (Ed.) J. van Ryzin, New York: Academic Press, 17-44.*

MACQUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 231-297.

MCLACHLAN, G.J. (1982), "The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis," In *Handbook of Statistics*, Vol. 2, (Eds.) P. R. Krishnaiah and L. N. Kanal, Amsterdam: North-Holland Publishing Company, 199-208.

MORGAN, B.J.T.)1981), "Three Applications of Methods of Cluster-Analysis," *Statistician, 30*, 205-223.

MUNGOMERY, V.E., SHORTER, R., and BYTH, D.E. (1974), "Genotype X Environment Interactions and Environmental Adaption. I. Pattern Analysis - Application to Soya Bean Populations," *Australian Journal of Agricultural Research, 25*, 59-72.

RAMSAY, J.O. (1982), "Some Statistical Approaches to Multidimensional Scaling Data," *Journal of the Royal Statistical Society, A 145*, 285-312.

SHEPARD, R.N. (1962a), "Analysis of PRoximities: Multidimensional Scaling with an Unknown Distance Function. I," *Psychometrika, 27*, 125-140.

SHEPARD, R.N. (1962b), "Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II," *Psychometrika, 27*, 219-246.

SHORTER, R., BYTH, D.E., and MUNGOMERY, V.E. (1977), "Genotype X Environment Interactions and Environmental Adaptation. II. Assessment of Environmental Contributions," *Australian Journal of Agricultural Research, 28*, 223-235.

WHITMORE, R.C., and HARNER, E.J. (1980), "Analyses of Multivariately Determined Community Matrices Using Cluster Analysis and Multidimensional Scaling," *Biometrical Journal, 22*, 715-723.

WOLFE, J.H. (1971), "A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions," *Naval Personnel and Training Research Laboratory, Technical Bulletin STB 72-2*, San Diego, California.