Theor Appl Genet (1985) 70:585–589

Multi-criteria clustering in genotype-environment interaction problems*

L. P. Lefkovitch

Agriculture Canada, Engineering and Statistical Research Institute, Central Experimental Farm, Ottawa, Ontario K1A OC6, Canada

Accepted January 19, 1985
Communicated by J. MacKey

Summary. Significant genotype-environment interactions in an ANOVA can be found for a number of reasons: one is the differences in the among-environments variances for each genotype, another is the differences in the ordering of the environments by each genotype. Using conditional clustering, groups may be formed in which the means, variances and patterns are used simultaneously but separately to decide on group homogeneity.

Key words: Grouping genotypes – Conditional clustering – Multiple criteria

Introduction

According to Lin (1982), two aspects of the data structure in the context of genotype (G) by environment (E) experimentation are of special importance, namely, the 'level' aspect, represented by the marginal means, and the 'shape' aspect, represented by the differential responses of individuals to one factor at different levels of the other. He, and others (see references cited by Lin) argue that the genotypes (or environments) should be grouped so that there will be no significant GE interaction within groups. Rarely, however, are the reasons for the presence of such an interaction considered, which can include the range of values which different genotypes may show among environments, and, quite separately, the pattern of highs and lows which they may show even if the ranges are the same. This distinction can be illustrated by reference to Table 1, in which the (hypothetical) mean yields of a replicated trial are given for four varieties grown in five locations. Considering all five locations, if only varieties I and IV had been grown, an ANOVA would not suggest a significant interaction, but would do so for each of the other pairs. For (I, II) the interaction can be explained by the different among-locations variances for them, for (I, III) by the different ordering of the locations with respect to the mean yields, while for (II, III) it is by both reasons.

The purpose of this paper is to draw attention to the fact that the among-environments variance for each genotype conveys information relevant to their grouping which is separate from the pattern of highs and lows, and also that clustering is possible in which there are several independent measures of relationship without the need to combine the latter. Other pertinent literature on the problems of grouping in the present context is cited by Lin (1982) and does not need to be reviewed again.

Table 1. Hypothetical mean yields of four varieties grown in five locations

Varieties	Loacations					Mean	Among-locations variance
	a	b	c	d	e		
I	1	3	5	7	9	5	10
II	3	4	5	6	7	5	2½
III	9	5	1	7	3	5	10
IV	11	13	15	17	19	15	10
Mean	6	6¼	6½	9¼	9½	7½	
Among-varieties variance	$22\frac{2}{3}$	$20\frac{11}{12}$	$35\frac{2}{3}$	$26\frac{11}{12}$	$46\frac{1}{3}$		26.58

* Contribution No. I-685 from the Engineering and Statistical Research Institute

Method

Let x_{ik} be the observed mean response of the ith genotype in the kth environment, and m_i and s_i^2 the mean response and among-environments variance respectively (the m_i and s_i^2 need not be related). If the m_i are equal in a probabilistic sense, or fall into a number of subsets in which they are equal, it seems reasonable further to group the genotypes belonging to each of these subsets on the s_i^2, since, for example, the yield of the varieties belonging to a subset having the smallest among-locations variance tends to be independent of the differences in the locations. Assume for the moment that the genotypes have been grouped into subsets for which the means and among-environments variances are homogeneous; if each such group contains more than one genotype, and exhibits a significant GE interaction, this can only be due to "shape". Thus the problem is to determine if interesting subsets of genotypes exist in groups exhibiting a GE interaction for which it is assumed that all have identical means and among-environments variances.

With the assumptions just made, all that remains is the pattern of the responses across the environments, which being multivariate, does not lead to an ordering or even a partial ordering. The motivation for what follows is that if these patterns are the same for some subset of the genotypes of interest, then it is reasonable to assume that those genotypes showing this pattern are equivalent, and given that there is no other relevant (external) factor, any member of a subset can be used for any other. Thus the problem becomes that of finding subsets of the genotypes for which the patterns within a subset are very much more alike each other than they are to the patterns shown by the members of any other subset. The first problem, therefore, is to describe the difference in the patterns in some way which is independent of the m_i and s_i^2.

Since the m_i are assumed to be (probabilistically) identical and to have a common s_i^2, the x_{ik} are first translated so that the variety mean becomes zero, and then normalized to have unit variance; this is achieved if x_{ik} is replaced by v_{ik} defined as

$$v_{ik}=(x_{ik}-m_i)/\parallel x_{ik}-m_i\parallel .$$

If v_i denotes the vector $\{v_{ik}\}$, the Euclidean distance between v_i and v_j, denoted by d_{ij} is given by

$$\begin{aligned} d_{ij}^2 &= (v_i-v_j)'(v_i-v_j) \\ &= 2\,(1-v_i'v_j) \\ &= 2\,(1-\cos\theta_{ij}) \end{aligned}$$

where θ_{ij} is the angle between the vectors v_i and v_j. It can be seen that d_{ij} is the linear distance between the ends of unit vectors, which, together with θ_{ij}, is independent not only of the means but also of the among-environments variance. Furthermore, even if the genotypes truly have unequal variances and/or means, these values focus just on the patterns, and exclude components arising from the means and variances. If the squared distance among genotypes had been computed without normalization to unit variance, the value obtained can be written as $s_i^2+s_j^2-s_is_j\cos\theta_{ij}$ which is twice the value obtained by Lin (1982). This value has the form of the variance of a difference, with expectation twice the error variance when the null model for GE interaction is true.

In essence, therefore, the complete grouping procedure uses three criteria: (1) the formation of subsets of genotypes in each of which the means are not significantly different; (2) for each of the subsets satisfying the first criterion, further division into subsets in each of which the among-environments variances are homogeneous; and (3) for each subset satisfying the first two criteria, further division into subsets in each of which the pattern of highs and lows are the same. As will be shown below, however, these three criteria do not require three distinct clustering steps.

The assumption that the variability of each genotype at each environment is independent of the mean, x_{ik} (this variability is not available for the data of Table 2) does not imply that the s_i^2 and m_i are independent. If they are related, then using the means independently from the among-environments variance as if these two components are independent may be misleading. These measures can be combined for simultaneous use in a clustering context by replacing them by the Fréchet distance (Dowson and Landau 1982), δ_{ij} defined as

$$\delta_{ij}^2=(m_i-m_j)^2+(s_i-s_j)^2$$

which will be zero if and only if the genotypes have identical means and variances. Using the Fréchet distances in a clustering context will then form groups which will be alike with respect both to the means and the among-environments variances.

At this stage, therefore, the data have been replaced either by three sets of values, namely, the means, variances and pattern distances, or by two sets, namely, the Fréchet and pattern distances. These are independent of each other, and at least one set is not uni-dimensional; thus it becomes necessary to employ a clustering procedure if groups are to be formed empirically. While each measure of distance may be processed by any greedy-type algorithm, such as employed by Lin (1982), the objective is really to form groups which are alike simultaneously on all sets of distances. One possibility for this is to find some compromise measure of distance (Lefkovitch 1978), but this has its own problems, and then employ the hierarchical clustering algorithm used by Lin, described originally by Sokal and Michener (1958), or replacing it by some other such procedure which is computationally faster and which exhibits fewer pathologies (Fisher and Van Ness 1971). Unfortunately, hierarchical procedures do not lend themselves to simultaneous clustering on several criteria in a simple and natural manner, and so it is of value to consider a clustering method which can use more than one set of distances. One such procedure, conditional clustering (Lefkovitch 1980, 1982), which produces subsets directly (and not dendrograms), re-expressed for the present circumstances, is:

```
for i = 2 . . . n
    for j = 1 . . . i-1
        group = {object i, object j} ;
        if the group is acceptable then
        label: for k = 1 . . . n, consult the 'oracle' to decide
                    whether k should be a member or not;
                    end k;
        if the group has been changed on the last pass
          through the for k loop, go to label;
        store the group for later processing;
    end j;
end i;
```

The generated groups are then examined in a number of ways; for example, if one of the groups consists of all genotypes, it can be discarded. From these considerations, it follows that there is no point in considering as an initial pair those objects which are least alike on any of the criteria, since a final group containing them will contain all n objects. It can be shown that the only initial pairs of objects which need be considered are those which are adjacent on the relative neighbourhood graph (Toussaint 1980) for each measure of distance under consideration. Details on the criteria for the 'oracle' are given by Lefkovitch (1982); essentially, it makes decisions based on

a comparison between the maximum distance amongst the current members with the average distance a candidate has to members of the group (note that measures of pairwise distance for which there is a theoretical upper bound of z are to be replaced by $-\log(z - d_{ij})$). Thus the oracle decides that object k should be adjoined to the group if it is sufficiently alike the current members on all criteria (n.b. this may also include the among-years variance). The procedures for obtaining the maximal joint probability solution from the generated subsets are essentially that of least-cost set covering applied to the minimum cross-entropy probabilities of each subset, and are described by Lefkovitch (1982).

There is no difficulty in extending conditional clustering to achieve groups homogeneous simultaneously for both genotypes and environments. For the moment, consider environments for which it is not possible to specify their mutual geographical proximities; then using the environments as the objects of the basic alogirthm, the oracle decides membership of an environment group on the basis of come set of distances (e.g. the mean yield at the environment, the among-genotypes variances, the pattern of genotypes at an environment) based on all or on a subset of the genotypes. If the mutual proximity of the environments is known, and can be represented by a Gabriel graph (Toussaint 1980), then it is not necessary to consider all others as possible candidates, since the initial pairs can be confined to those environments which are adjacent on this graph, and candidates for admission can be confined to those which form a connected subgraph with the current members (Lefkovitch 1980).

Discussion

This paper has distinguished three components of the responses of a genotype to a range of environments which are of interest, namely, the genotype mean ignoring environments, the among environments variance, and the pattern of responses among the environments, and has described a clustering procedure in which these can be used simultaneously without the need to combine them. Furthermore, the incorporation of year-to-year variability does not require anything new other than to include an appropriate measure of this as part of the decision process performed by the 'oracle' in the algorithm. This variability can also be included in the Fréchet distances; if y_i^2 represents the among-years variance for the ith genotype, then

$$\delta_{ij}^2 = (m_i - m_j)^2 + (s_i - s_j)^2 + (y_i - y_j)^2$$

is also a Fréchet distance; the definition can be extended by the inclusion of further terms in comparable units. Furthermore, the definition of the pattern vectors can also be extended to include further elements to represent the additional data.

The conditional clustering algorithm can also be used for the grouping of any sets of objects based on more than one set of relationships e.g. a grouping of the environments. A procedure for grouping genotypes and environments simultaneously is also possible; that suggested above, which can be regarded as a marriage between the two-way alternating strategy of Hartigan (1972) with that of conditional clustering, is obvious and does not require further comment.

Table 2. Variety means (based on three replications and two years data) cited from Yates and Cochran (1938)

Variety	Locations						Mean	Variance
	1	2	3	4	5	6		
'Manchuria'	161.7	247.0	185.4	218.7	165.3	154.6	188.8	1,349.82
'Svansota'	187.7	257.5	182.4	183.3	138.9	143.8	182.3	1,810.20
'Velvet'	200.1	262.9	194.9	220.2	165.8	146.3	198.4	1,685.55
'Trebi'	196.9	339.2	271.2	266.3	151.2	193.6	236.4	4,664.80
'Peatland'	182.5	253.8	219.2	200.5	184.4	190.1	205.1	751.18
Mean	185.8	272.1	210.6	217.8	161.1	165.7	212.1	
Variance	230.8	1,441.3	1,335.8	962.1	293.4	588.2		1,689.98

Table 3. Pattern vectors for the varieties in Table 2

		Locations					
		1	2	3	4	5	6
'Manchuria'	(1)	−0.3299	0.7084	−0.0414	0.3640	−0.2861	−0.4163
'Svansota'	(2)	0.0568	0.7904	0.0011	0.0105	−0.4562	−0.4046
'Velvet'	(3)	0.0185	0.7025	−0.0425	0.2374	−0.3550	−0.5674
'Trebi'	(4)	−0.2586	0.6731	0.2279	0.1958	−0.5579	−0.2802
'Peatland'	(5)	−0.3688	0.7946	0.2301	−0.0751	−0.3378	−0.2448

Numerical example

To illustrate the arguments just presented, the data used by Lin (1982) will be reconsidered the data are given in Table 2, together with the means and among-locations variances. The pattern vectors are given in Table 3, the squared distances and angles in Table 4, Lin's estimated squared distances (which equate the means but which have not been based on equal among-locations variances) in Table 5, and the squared Fréchet distances in Table 6. The differences are striking; while the largest distance in Table 5 is between 'Trebi' and 'Peatland', it is ranked fifth (out of 10) in Table 4. Considering just the variety means and variances, those for 'Trebi' and 'Peatland' suggest that they are somewhat different from each other and from the other three without any consideration whatsoever of the pattern of responses across locations, which is essentially the grouping obtained by Lin.

This grouping is confirmed by an examination of the Fréchet squared distances, in which it is apparent that 'Trebi' is unlike the other four, and also that 'Peatland' is also somewhat different from the others, although less so. It is also apparent from Table 4 that grouping on pattern alone would give a different arrangement of the varieties. This difference is confirmed by the results of a conditional clustering (Lefkovitch 1982) of each set alone of the distances which are the square roots of the values in Tables 4–6. Table 7, where a summary of the results is given, shows that pattern without variance associates two varieties which show the highest and lowest among-locations variance (Table 2). It is easy to see that the grouping obtained, corresponding to Tables 5 and 6, can be inferred almost by inspection of the among-locations variances in Table 2. A simultaneous conditional clustering using the Fréchet distances of Table 6 and the pattern distances of Table 4 yielded 4 distinct subsets (Table 7 d) in which each of 'Trebi' and 'Peatland' are single-object subsets, and two overlapping subsets, namely, {'Manchuria', 'Velvet'} and {'Svansota', 'Velvet'}, forming a three-object muster {'Manchuria', 'Velvet', 'Svansota'}. This solution, hinted at by combining the separate analyses of the data of Tables 4 and 6 separately, seems to be the most suitable for these

Table 4. Squared distances among the varieties based on the pattern vectors (below diagonal) and angles in radians among them (above diagonal)

(1)	–	0.5663	0.4048	0.4506	0.5618
(2)	0.3122	–	0.3164	0.4785	0.5361
(3)	0.1632	0.0993	–	0.5317	0.6719
(4)	0.1996	0.2247	0.2761	–	0.3897
(5)	0.3074	0.2806	0.4348	0.1500	–

Table 5. Squared distance equating means but without equating among-locations variances (from Lin 1982)

(1)	–				
(2)	260.7	–			
(3)	133.8	88.5	–		
(4)	748.3	658.3	755.0	–	
(5)	198.3	278.1	336.5	976.4	–

Table 6. Squared Fréchet distance based on variety means and among-locations variances

(1)	–				
(2)	75.17	–			
(3)	110.00	261.50	–		
(4)	3,268.26	3,598.89	2,196.82	–	
(5)	353.66	748.37	229.96	2,664.11	–

Table 7. Summary of results obtained from conditional clustering of the data in Tables 4–6

(a) From Table 6	
Generated subsets	Probability[a]
(1) {1, 2, 3, 4}	1.0
(2) {4}	1.0
(b) From Table 5	
(1) {1, 5}	0.25
(2) {1, 2, 3, 5}	0.5
(3) {2, 3}	0.25
(4) {4}	1.0
optimal solution[b] consists of subsets (2) and (4)	
(c) From Table 4	
(1) {1, 3}	0.5
(2) {1, 4}	0.5
(3) {4, 5}	1.0
(4) {2, 3}	1.0
optimal solutions[b] (i) subsets (1), (3), (4). (ii) subsets (2), (3), (4).	
(both optimal solutions generate the same musters, namely, {1, 2, 3}, {4, 5}	
(d) From Tables 4 and 6 simultaneously	
(1) {1, 3}	1.0
(2) {2, 3}	1.0
(3) {4}	1.0
(4) {5}	1.0
All four subsets required to cover the varieties There are three musters: {1, 2, 3}, {4}, {5}	

[a] Maximum entropy estimates; see Lefkovitch (1982)
[b] Solutions which maximise the joint probability of the chosen subsets

Table 8. Analysis of variance for the data of Table 2 with and without variety groupings suggested by the simultaneous clustering based on the Fréchet and pattern distances, together with a decomposition of the residuals

Source	DF	SS %
Varieties (V)	4	17.15
Groupings (G)	2	15.88
Within group 1	2	1.27
Locations (L)	5	68.53
Residual V.L	20	14.31
(a) Without grouping		
Singular vectors		
1st	4	9.37
2nd	5	2.96
3rd	5	1.86
4th	5	0.12
(b) With grouping		
G.L	10	11.72
W.G.L	10	2.59
Singular vectors		
1st	5	2.13
2nd	5	0.46
Total	29	100 (actual value 61927.8)

data. Table 8 gives an analysis of variance for the three-group arrangement from which it can be seen that the variety groupings absorb over 90% of the sums of squares estimated for differences among the varieties. The sum of squares associated with the entry W.G.L in Table 8 was further analyzed by a decomposition of the 5×6 array of residuals (Snee 1982); the rank of this array, which is confined to group 1 containing three varieties, cannot exceed 2. The actual sum of squares for W.G.L is 1,606.85; the squared singular values are 1,321.12 and 285.73; the singular vector associated with the larger of these is the contrast [0.754–0.648–0.106] corresponding to 'Manchuria', 'Svansota' and 'Velvet', which suggests that 'Velvet' is intermediate between the other two (note that conditional clustering placed Velvet in the intersection of the multi-object subsets) and also that 'Manchuria' is perhaps somewhat different from the other pair. The 'Bartlett' test for the equality of these two squared singular values gives a value of 4.16, which as a chi-square with 5 df is not significant. By contrast, a similar decomposition of the residuals without any grouping of the varieties (Table 8) gave a value of 56.57 which as a chi-square with 9 df clearly indicates heterogeneity in the interaction structure.

Acknowledgements. This paper has benefited considerably from discussions with Drs. C. S. Lin and B. K. Thompson, and particularly, from comments made on an early draft by Mrs. P. M. Morse. It should be noted, however, that the opinions expressed in the paper are not necessarily supported by them.

References

Dowson DC, Landau BV (1982) The Fréchet distance between multivariate normal distributions. J Multivar Anal 12:450–455

Fisher L, Van Ness JW (1971) Admissible clustering procedures. Biometrika 58:679–691

Hartigan JA (1972) Direct clustering of a data matrix. J Am Stat Assoc 67:123–129

Lefkovitch LP (1978) Consensus coordinates from qualitative and quantitative attributes. Biom J 20:679–691

Lefkovitch LP (1980) Conditional clustering. Biometrics 36: 43–48

Lefkovitch LP (1982) Conditional clusters, musters, and probability. Math Biosci 60:207–234

Lin CS (1982) Grouping genotypes by a cluster method directly related to genotype-environment interaction mean square. Theor Appl Genet 62:277–280

Snee RD (1982) Nonadditivity in a two-way classification: is it interaction or nonhomogeneous variance. J Am Stat Assoc 77:515–519

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci, Bull 38:1409–1438

Toussaint GT (1980) The relative neighbourhood graph of a finite planar graph. Pattern Recognition 12:261–268

Yates F, Cochran WG (1938) The analysis of groups of experiments. J Agric Sci 28:556–580