GROUPING OF VARIABLES WHEN
UNITS HAVE A NATURAL ORDER

by

S.R. Esterby

SUMMARY

The characterization of change in a multivariate response with naturally ordered units is considered. The methods commonly used, such as clustering or plotting in a reduced dimension, either ignore the natural order or assume step changes or both. The proposed method consists of grouping those response variables which have similar forms of change over the sequence of observation. Two procedures, one using least squares cubic splines and the other, using a similarity measure based on common runs up or down of smoothed values, are considered. The final characterization consists of sets of response variables and the form of change for each set.

KEY WORDS: Multivariate response; ordered observations; least squares cubic splines; cluster analysis.

RESUMÉ

On étudie la caractérisation du changement déterminé par la réponse multivariable naturellement ordonnée. Les méthodes habituellement utilisées, par exemple l'application de la théorie des grappes ou le traçage des courbes dans des dimensions réduites, sont insuffisantes soit parce qu'elles ignorent l'ordre naturel ou parce qu'elles supposent des changements par étape, ou pour ces deux raisons. La méthode proposée est basée sur le regroupement des variables de la réponse qui présentent des profils semblables de changement pour une séquence donnée d'observations. Deux méthodes, l'une basée sur l'utilisation de la fonction spline (méthodes des moindres carrés) et l'autre, utilisant une mesure de la ressemblance basée sur des essais communs effectués avec des valeurs normalisées, sont étudiées. La caractérisation finale est constituée d'ensembles de variables de réponses et du type de changement pour chaque ensemble.

1. INTRODUCTION

Consider a situation where a sequence of measurements on three or more response variables have been made and the objective is to characterize the change in these variables over the sequence. Many such multivariate data sets occur in the environmental sciences. The methods of data analysis usually include either a variable reduction technique or a unit grouping technique or both. Thus, for example, the scores on the first few principal components could be examined in the order of the sequence, or the sequence could be divided into groups using a method of cluster analysis. The variable reduction techniques such as principal components analysis and most cluster analysis methods do not take the order of observation into account explicitly. Constrained clustering methods (Gordon, 1980) do account for this natural order but characterize the change as that of a change from one constant level to another. Situations arise where the form of change is better characterized by a smooth or piecewise smooth curve. This paper considers ways of characterizing such data sets by grouping response variables which have similar curves.

The response variables could be different physical variables or the same physical variable with some other characteristic distinguishing the different response variables. Herein, the result of applying a smoothing procedure to the data will be called a curve, and the similarity of curves that will be considered is that the curves

increase and decrease together. This can be examined visually for variables which have different units or different scales by plotting each variable on axes of the same size and fixing minimum and maximum values of each response variable at the same point on the ordinate in each plot. The plots can be overlaid to see which are approximately the same. To mimic this in a mathematical procedure, the response variables may need to be rescaled so that each has minimum and maximum values of −C and C, where C=1 is a convenient choice. In cases where it is important to know if the magnitudes are the same for variables in the same units, variables would not be rescaled.

Suppose $(x_1, \underline{y}_1)$, $(x_2, \underline{y}_2)$, ..., $(x_n, \underline{y}_n)$ are the n observations on the p response variables, which have been rescaled if appropriate, where the subscript i gives the position in the sequence, and x denotes either the index i or the value of a variable such as time, which determines the order. Then the change in the jth response variable over the sequence is given by

$$y_{ij} = f_j(x_i) + \epsilon_{ij} \qquad (\text{for } i=1,2,\ldots,n) \qquad (1)$$

where f is some smooth or piece-wise smooth function and ϵ is the error random variable. The problem is to determine which of the f_j are the same. The general approach will be to estimate $\hat{f}_j$ for j=1,2, ..., p and then either use a step-wise approach to determine which variables can be fitted by the same function or a hierarchical

clustering method in which similarities between all possible combinations are calculated.

2.   REGRESSION

To put the present case in a framework analogous to partial F-tests in regression (Draper and Smith, 1981, section 2.9), the form of f is fixed and one function, say f_k, is fitted to response variables j and j', as well as fitting f_j and $f_{j'}$. The the extra variation, due to fitting fewer parameters in the single function f_k, is assessed. It will be assumed that the observations are independent but the response variables are not necessarily independent. The well known results for univariate (e.g., Draper and Smith, 1981) and multivariate (e.g., Kendall, Stuart and Ord, 1983) linear models will be used.

Polynomials

If the function can be modelled as a polynomial in x of degree q, (1) is given by

$$y_{ij} = \beta_{jo} + \beta_{j1} x_1 + \beta_{j2} x_1^2 + \ldots + \beta_{jq} x_1^q + \epsilon_{ij}$$

$$\text{(for } i=1,2,\ldots,n) \qquad (2)$$

where $j=1,2,\ldots,p$. To determine if the response of variables j and j' can be characterized by the same polynomial,

$$y_{ik} = \beta_{k0} + \beta_{k1} x_1 + \beta_{k2} x_1^2 + \ldots + \beta_{kq} x_1^q + \epsilon_{ik} \qquad (3)$$

for $k=j$ or j' and $i=1,2,\ldots,n$, is also fitted. Under the assumptions that the ϵ_{ij} are normally distributed with $E(\epsilon_{ij}) = 0$ and that the response variables are independent, the increase in residual sums of squares from model (3) with $(q+1)$ parameters relative to model (2) with $2(q+1)$ parameters is tested using an F statistic. If the response variables are dependent, the likelihood ratio test for the two models is used.

Cubic splines

For environmental data, single polynomials are often not adequate. Further, the number of response variables may be large and the forms of the curves quite different. To develop a procedure that will provide reasonable results without user interaction at the fitting

stage, a more general smoothing procedure is needed. Least squares cubic splines with fixed knots is one such procedure and has the advantage that it is equivalent to restricted least squares (Buse and Lim, 1977).

Using cubic splines, (1) takes the form

$$y_{ij} = \beta_{j0\ell} + \beta_{j1\ell} x_i + \beta_{j2\ell} x_i^2 + \beta_{j3\ell} x_i^3 + \varepsilon_{ij}$$

$$\text{for } d_{\ell-1} \leq x_i \leq d_\ell \text{ and } i=1,2,\ldots,n \tag{4}$$

where the (m+1) knots $d_0, d_1, \ldots, d_m$ divide the domain of x into m intervals. In addition, the continuity restrictions on the cubic polynomials and their first and second derivatives at the interior knots provide the full set of equations and a parameter space of dimension m+3 for each j. The estimation could be performed by restricted least squares, or upon imposition of two additional restrictions, by least squares cubic splines (see Buse and Lim, 1977, for details). Analogous to the case of single polynomials, the model for variables j and j' together is

$$y_{ik} = \beta_{k0\ell} + \beta_{k1\ell} x_i + \beta_{k2\ell} x_i^2 + \beta_{k3\ell} x_i^3 + \varepsilon_{ij}$$

$$\text{for } k=j \text{ or } j', \ d_{\ell-1} \leq x_i \leq d_\ell \text{ and } i=1,2,\ldots,n \tag{5}$$

subject to the restrictions given above. Assuming normality, the tests based on the change in the residual sum of squares or the

likelihood ratio are used for independent or dependent response variables, respectively. In general, to test that a set of q variables, with values of j given by the set J_q, can be fitted by the same function, (5) is defined for k given by J_q.

This can be programmed easily using a least squares cubic spline procedure such as the IMSL subroutine ICSFKU (IMSL, 1982), which was used for the examples here. Let $\hat{y}_{ij}$ be the value of the cubic spline for variable j calculated at x_i and $\hat{\epsilon}_{ij} = y_{ij} - \hat{y}_{ij}$ be the corresponding residual. If J_q gives the variables to be combined, the matrix of residuals is $R = (\hat{\epsilon}_{ij})$, of dimension nxq. Let R_u denote the matrix calculated under (4) and R_c that under (5), where both matrices are of dimension nxq. Thus to test if the variables identified in J_q can be fitted by the same spline, the hypothesis is that $f_j = f_k$ for j an element of J_q, and the test statistic is

$$F = (qn - q(m+1)) \, (RSS_c - RSS_u)/(q-1)(m+1) \, RSS_u$$

where RSS_c is obtained from R_c and RSS_u from R_u as

$$\sum_{j \in J_q} \sum_{i=1}^{n} \hat{\epsilon}_{ij}^2.$$

The statistic follows an F distribution with $(q-1)(m+1)$ and $qn - q(m+1)$ degrees of freedom. The statistic in the case of dependent variables is

$$\lambda = -n \log_e |\hat{\Sigma}_u| / |\hat{\Sigma}_c|$$

where $\hat{\Sigma}_u = (1/n) R'_u R_u$ and $\hat{\Sigma}_c = (1/n) R'_c R_c$, both of dimension $q \times q$. The distribution of λ can be approximated by the chi-square distribution with $q(m+1)$ degrees of freedom (Kendall, Stuart and Ord, 1983).

A selection procedure

To avoid calculating the splines for all possible combinations of two or more variables some procedure is needed for determining the order in which combinations are to be fitted by one spline. The method used in this paper is as follows:

1. Fit splines to all p variables and calculate the fitted values $\hat{y}_{ij}$.

2. Calculate the correlation coefficient between vectors of fitted values for all possible pairs of variables, and exclude any pairs with correlation less than some arbitrary value, r_c, where $r_c \geqslant 0$.

3. Order the remaining pairs of variables to correspond to decreasing value of the correlation coefficient. Choose the pair with the largest correlation coefficient.

4. Test whether this pair of variables can be fitted by the same spline. If not, go to step 8. If so, go to step 5.

5. Calculate the correlation coefficient between the fitted values $\hat{y}_{ik}$, obtained from the spline of the combined variables, and all variables not appearing in this combination or a previous combination found to be fitted adequately by one spline. Choose the variable with the largest correlation coefficient.

6. Calculate the spline for the enlarged combination. Test whether this combination can be fitted by the same spline.

7. If so, and all columns have not been tested, go to 5. If not and all the columns have not been tested, go to step 8. Otherwise, stop.

8. From the remaining pairs, for which neither member has appeared in a previous combination found to be fitted adequately by a single spline, choose the pair with the next highest correlation. If such a pair is found, go to step 4. Otherwise stop.

For the examples, $r_c=0$ was used at step 2. This prevented the value of r_c from controlling which variables would be tested, except to exclude those which were negatively correlated.

Choosing the form of f_j

Like all methods of data analysis, some preliminary examination of the data and subsequent assessment of the analyses are required. The critical initial question is the choice of the form of f_j so that it will be adequate for all variables. If cubic splines are used, the location and number of knots need to be specified. How this is done will depend upon the situation, since there may be information available about how much variation to smooth out and where knots should be placed. In the absence of such information, plots of the variables versus x will help in the choice. If some of continuity assumptions are inappropriate, the estimation can be done by restricted least squares.

3. CLUSTERING

As in the previous section, the idea of smoothing out the excess variation and measuring the similarity between the smoothed curves is applied. Measures of association between variables, commonly used for clustering, such as a correlation coefficient, are not suitable for the present purpose since they cannot distinguish between runs of similarity in curves and disjoint points of similarity. The measure of similarity defined here uses runs of points where both smoothed

curves increase, decrease or remain constant, and weights the contribution of a run so that runs corresponding to large peaks or troughs contribute more to the similarity coefficient than do runs of small peaks or troughs.

Let $\bar{y}_{ij}$ denote the value of the running mean of length L calculated for variable j at index i of the sequence. To obtain the similarity between the running means of variables j and ℓ, the vector, $\underline{m}$, of matches between the first differences within the two columns is given by $\underline{m} = (m_1, \ldots, m_{n-1})'$ where

$$m_i = \begin{cases} 1 \text{ if sgn } (\bar{y}_{i+1,j} - \bar{y}_{i,j}) = \text{sgn } (\bar{y}_{i+1,\ell} - \bar{y}_{i,\ell}) \\ 0 \text{ otherwise} \end{cases}$$

and

$$\text{sgn}(z) = \begin{cases} -1 & z<0 \\ 0 & z=0 \\ 1 & z>0 \end{cases}$$

The ith element of $\underline{m}$ is based on the ith and (i+1)st elements of $\underline{y}_j$ and $\underline{y}_\ell$. Thus to define the weighting factor, let

$$\tilde{y}_{ij} = 1/2 \; (\bar{y}_{i+1,j} + \bar{y}_{i,j})$$

for $i=1,2,\ldots,n-1$ and $j=1,2,\ldots,p$. Then the ith element of the vector of weights, $\underline{w}$, is given by

$$w_i = | r_{ij} + r_{i\ell} |$$

where r_{ij} = signed rank of $\dot{y}_{ij}$ relative to the median rank for variable j, and $r_{i\ell}$ is similarly defined. Thus, for $(n-1)=11$, r_{ij} and $r_{i\ell}$ will range from -5 to $+5$ with median rank of 0.

Runs of 1's in the vector m, hereafter referred to only as runs, indicate regions of agreement between variables j and ℓ. For the kth run, the contribution to the measure of similarity is

$$s_k = \sum_i m_i w_i$$

for i a member of the kth run. Thus $\sum\limits_{k=1}^{n_r} s_k = \sum\limits_{i=1}^{n-1} m_i w_i$ where n_r = number of runs. Two measures of similarity which can be defined from this are

$$S_1 = \sum_{k=1}^{n_r} \delta_k s_k$$

where

$$\delta_k = \begin{cases} 1 \text{ if the kth run is of length} \geq LR \\ 0 \text{ otherwise} \end{cases}$$

and

$$S_2 = s_{max}$$

where $s_{max} = s_k$ for k corresponding to the longest run of length $\geq$ LR. LR exceeds 2 but is arbitrary.

A clustering method that uses means of columns found to be similar at the previous step is the following:

1. Calculate the similarity, S (either S_1 or S_2), between all pairs of variables and denote this by $S_{j\ell}$ for variables j and ℓ.

2. The pair(s) with max value of $S_{j\ell}$ are combined and a new running mean calculated from means of the original data.

3. The similarity matrix ($S_{j\ell}$) is now calculated between the reduced number of vectors of running means.

4. Steps 2 and 3 are repeated until all variables are in one group.

The above method, although hierarchical, since once a variable is entered into a cluster it does not move from that cluster, does not have the usual property that the similarity measure decreases monotonically as the number of clusters decreases. An alternative, which permits a conventional dendrogram to be constructed, is to perform a single or complete linkage algorithm on the original similarity matrix.

The nature of the data set and the objectives of the analysis will determine the values of L and LR. The length of the running mean, L, should be large enough to remove variability that is not part of the function being characterized by the curve. For a given level of scatter, broad peaks will require larger L than narrow peaks. The

value of LR, typically small, will depend upon the length of the sequence and the level of variability.

EXAMPLES

Sediment cores are taken from lake bottoms and sampled at various intervals, with the samples then being subjected to biological or chemical analysis, in order to draw inferences about past environmental conditions. The resulting data sets are multivariate but have a natural order since increasing depth in the core corresponds to more distant times in the past. Further, the number of variables for each sample is usually large but the number of samples small to moderate. The following examples are from such data sets.

Example 1

The abundances of eight pollen types in a sediment core were studied for evidence of changes in land vegetation in the past (Delorme et al., 1984). This set has previously been examined individually and by multivariate methods (Esterby et al., 1986). For all eight profiles the form of change is curvilinear with abrupt changes occurring low in the core for four types and nearer the surface for

the remaining four. Segmented polynomials were fitted in the process of estimating the point of discontinuity using the method of Esterby and El-Shaarawi (1981) and from inspection of the plots it was noted that, of the 4 profiles with the discontinuity low in the core, the members within each of the two pairs, Artemisia-Betula and Pinus-Tsuga, are most similar. Of the other 4, only the pair Ambrosia-Acer have similarly shaped curves. Picea and Fagus were not similar to any others.

Using these two points of discontinuity for the interior knots, the cubic spline selection procedure described above was used on the eight pollen types, after scaling all types to be in the interval -1 to 1. Based on the F-test and a 5 percent significance level, the groups of similar curves were: 1) the four pollen types with a discontinuity low in the core plus Picea and 2) Ambrosia and Acer, with Fagus not similar to any. When the likelihood ratio test, with the same level of significance, was used, only the pairs Ambrosia-Acer and Artemisia-Betula were found similar. The splines based on the first grouping, with the corresponding data points, are shown in Figures 1a and b. In Figures 1c and d, the splines obtained for the individual columns are also shown for the Artemisia-Betula combination and in Figure 1e that of Fagus, to illustrate how different it is. See Figure 2 for the correspondence between data matrix column numbers and pollen names.

The iterative clustering method, using similarity measure S_1, L=7 and LR=3, produced results in agreement with Esterby et al. (1986) as can be seen from the dendrogram and plots of running means (Figure 2). In this case S_1 did decrease monotonically.

As will be seen, the value of L here is higher than used for the other examples. Initially L=3 and L=5 were tried, but unreasonable combinations were obtained, and this was not remedied by increasing LR. By superimposing plots of the running means, each scaled to be in the interval [-1,1], unreasonable combinations are immediately obvious. The higher value of L produced curves more in agreement with the generally broad peaks fitted by segmented polynomials.

Example 2

The 14 most abundant diatom species, from the same core as the pollen data used in Example 1, have also been analyzed previously (Esterby et al., 1986). Here, the spline grouping method, with interior knots and significance level the same as in Example 1, and the iterative clustering method, the same except for L=5, were applied to the data. The three sets of results (Table 1) are in good agreement in the early stages of grouping, those which are of most interest, except for the inclusion of variable 9 with variables 11 and 10 by the clustering method. Variable 9 has considerably more scatter

which has less effect in this latter method. A combination of running mean plots and cubic spline plots, which characterize the major shapes of the curves, for this data set, are given in Figure 3. The results are in general agreement with the subjective inspection of segmented or single polynomials (Esterby et al., 1986), but this has not been discussed here because the present procedures have produced simpler summaries.

Example 3

From the complete enumeration of fossil pollen in a sediment core (McAndrews, 1966), a reduced set of 20 relatively abundant types, based on the reduced set of Gordon (1982), have been analyzed by the present methods, as described in the previous two examples (L=5 here). McAndrews prepared a pollen diagram and, from inspection of this diagram, divided the core into regions characteristic of high pollen abundance for one or more particular pollen types. The present methods look for similarity over the entire length of the core and thus provide an analysis complementary to the zonation of the core.

The iterative clustering method divided the 20 profiles into three groups, with only the early clusters in each group, i.e., those clustered at $(\max S_1-S_1) \leq 263$ (see lower horizontal line in Figure 4) being similar over the entire profile (Figures 4 and 5). Note that

the similarity is not strictly decreasing in this case, with variable 9 entering the existing group which consists of variables 15 and 20, with similarity greater than that for 15 and 20 alone, as indicated by the arrow in the dendrogram. The spline grouping procedure yielded only four pairs of similar variables with overlap between the F and likelihood ratio tests for three of the four pairs and the differences being due to the general tendency of the likelihood ratio version to produce higher significance probabilities and to the fact that it is a stepwise procedure.

The divergence of the results from the clustering and the spline grouping methods comes from the greater smoothing and the dependence on the variability as well as curve shape, for the latter method. It is much more pronounced for this example than the previous two, and would require the analyst to decide which level of smoothing is portraying the important features of the curves. Inspection of the running mean plots, with the boundaries of the pollen zones as given by McAndrews delimited, suggests that the pollen analyst would need to retain as much variability as retained by the running means to be able to define the pollen zones.

DISCUSSION

The methods of smoothing and of comparing smoothed curves used here were developed for data such as used in the examples, but should be applicable to many types of data. Within the general idea of smoothing and comparing smoothed curves, many variations are possible. The method could also be used for pairs of variables (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), where now both x and y are vectors, for a suitable standardization of x's. For example, to know if the response curve takes the same form for fixed increments in dose starting at different base levels for each x_j, the base level would be subtracted. The spline procedure can handle vectors of different lengths and the clustering procedure, since it compares curves, could also be used if some sensible common values of x could be found at which to estimate the curves. Fixed knots have been used so that the results of linear least squares are available. Further, because the rigour of the methods is that of clustering, it seems unwarranted to introduce the complexity of estimating knots. In the event of a poor first set of knots, the procedure can be repeated with another set.

It is essential that the curves are plotted since both procedures require specifications from the user, the values of which determine how well the procedures perform. However, this is not a disadvantage since the procedures group curves and in general our objectives will include knowledge of the shapes of the curves. To determine the

extent of agglomeration that can be used and still retain only similar curves within a cluster, clusters prior to a relatively large drop in similarity can be chosen and the corresponding plots examined. This was illustrated in Example 3.

ACKNOWLEDGEMENTS

The plotting programs for the cubic splines and running means were prepared by Joanne Hodson.

REFERENCES

Buse, A. and Lim, L. (1977). Cubic splines as a special case of restricted least squares. Journal of the American Statistical Association, 72, 64-68.

Delorme, L.D., Esterby, S.R. and Duthie, H. (1984). Prehistoric trends in pH for Kejimkujik Lake, Nova Scotia. Int. Revue ges. Hydrobiol., 69, 41-55.

Draper, N.R. and Smith, H. (1981). Applied Regression Analysis. New York: Wiley.

Esterby, S.R. and El-Shaarawi, A.H. (1981). Inference about the point of change in a regression model. Applied Statistics, $\underline{30}$, 277-285.

Esterby, S.R., Delorme, L.D., Duthie, H. and Harper, N.S. (1986). Analysis of multiple depth profiles in sediment cores: an application to pollen and diatom data from lakes sensitive to acidification. Hydrobiologia, $\underline{141}$, 207-235.

Gordon, A.D. (1980). Methods of constrained classification. In: Analyse de Données et Informatique (ed. R. Tomassone), I.R.I.A., Le Chesnay.

Gordon, A.D. (1982). Numerical methods in Quaternary palaeoecology. V. Simultaneous graphical representation of the levels and taxa in a pollen diagram. Review of Palaeobotany and Palynology, 37, 155-183.

Kendall, M., Stuart, A. and Ord, J.K. (1983). The Advanced Theory of Statistics, Vol. 3. 4th Edn. New York: MacMillan.

McAndrews, J.H. (1966). Postglacial history of prairie, savanna and forest in Northwestern Minnesota. Memoirs of the Torrey Botanical Club, 22(2), 1-72.

FIGURE CAPTIONS

Figure 1. Graphical summary from the spline grouping procedure for Example 1.

Figure 2. Dendrogram and plots of data and running means for Example 1.

Figure 3. Data and splines based on the columns of data shown on the plot (a,b,c and d) and data and running means (e,f,g and h) for Example 2.

Figure 4. Dendrogram for Example 3.

Figure 5. Plots of running means only for three major clusters shown in Figure 4 (plots on the left), and plots of running means and data for corresponding subsets with high similarity (plots on the right) for Example 3.

TABLE 1: Grouping of Variables, Example 2

Cubic Spline Grouping Method

F-Test

Variable	s.l.
11 10	0.95
11 10 7	0.98
11 10 7 3	0.91
11 10 7 3 5	0.50
11 10 7 3 5 9	0.80
11 10 7 3 5 9 8	0.56
11 10 7 3 5 9 8 2	0.17
11 10 7 3 5 9 8 2 13	0.09
14 1	0.13
14 1 12	0.14

LR-Test

Variable	s.l.
11 10	0.76
11 10 7	0.72
11 10 7 3	0.67
11 10 7 3 5	0.07
11 10 7 3 5 9	0.13
8 2	0.06
12 4	0.06

Iterative Clustering Method

Variable	S_1
11 9	135
8 2	133
1 14	125
7 10	110
11 9 7 10	96
5 11 9 7 10	91
3 8 2 5 11 9 7 10	88
4 12	65
6 1 14	61
13 4 12	40
6 1 14 13 4 12	21
all variables	17

s.l. observed significance level
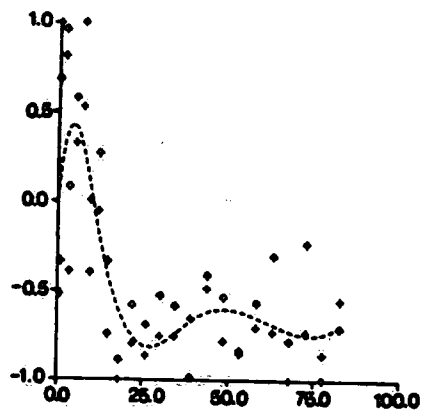
a) Data and combined spline, columns 3,6.
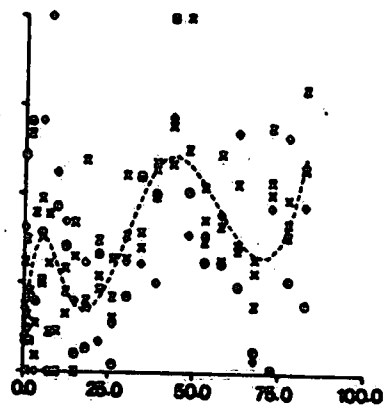
b) Data and combined splines, columns 7,4,5,1,8.
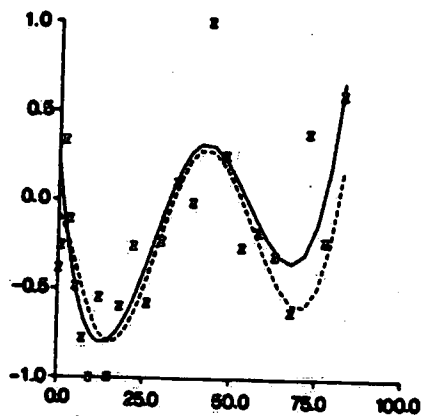
c) Combined spline, columns 1,8; data and spline column 8 (——)
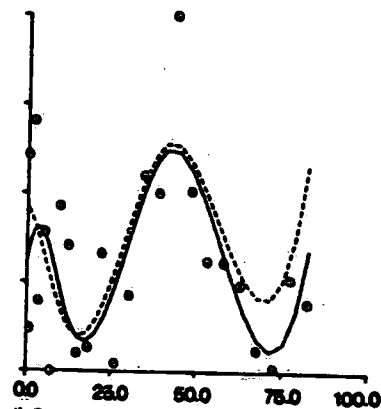
d) Combined spline columns 1,8; data and spline column 1 (——)
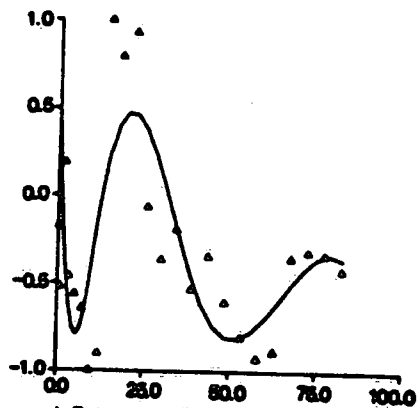
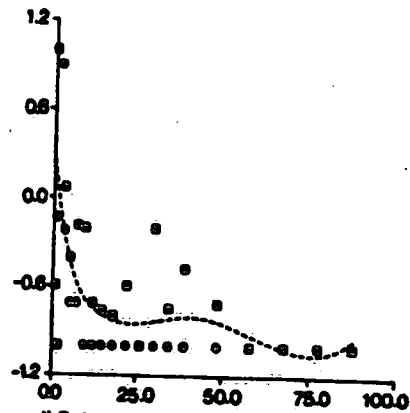e) Data and spline, column 2

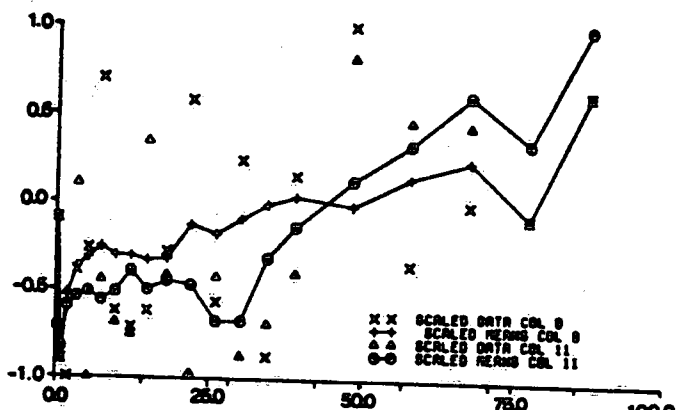Fig. 1
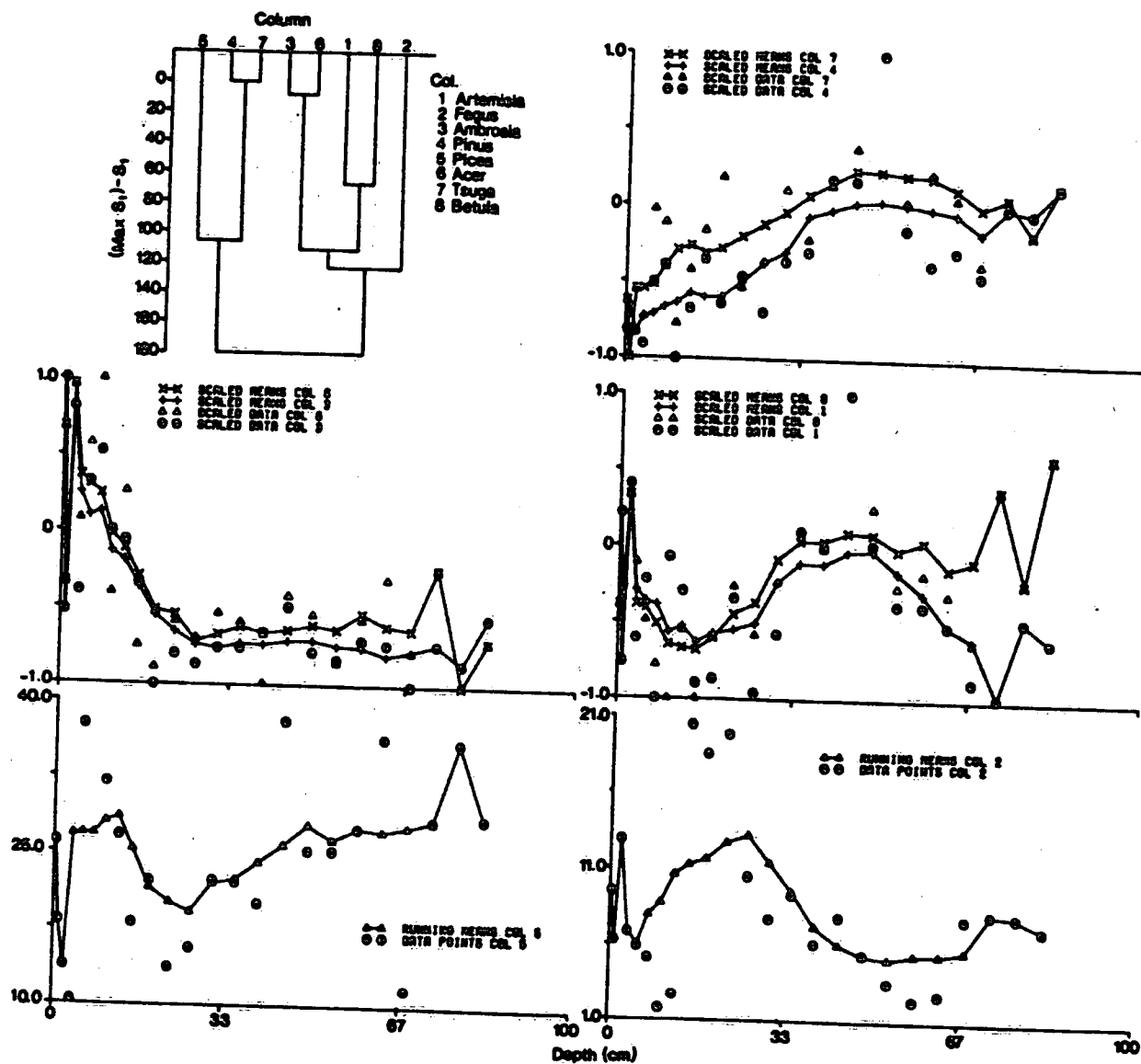
Fig. 2
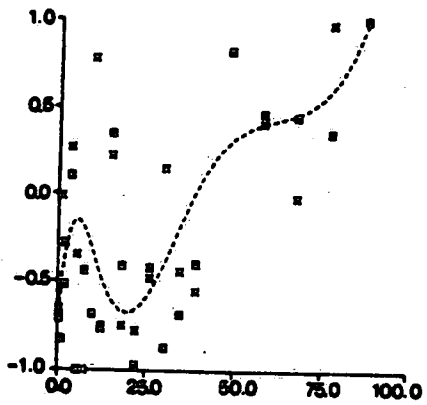
a) Data and combined spline, columns 11,10
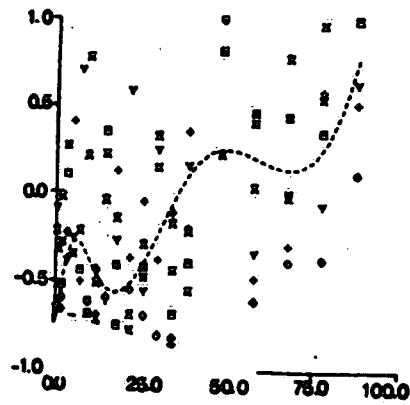
b) Data and combined spline, columns 11,10,7,3,5,8
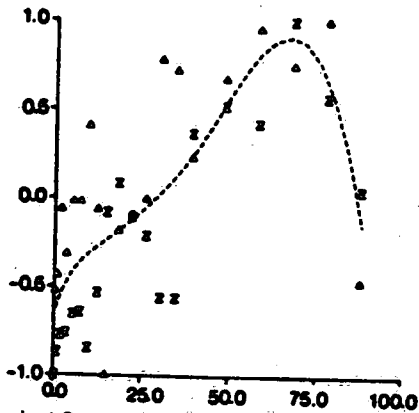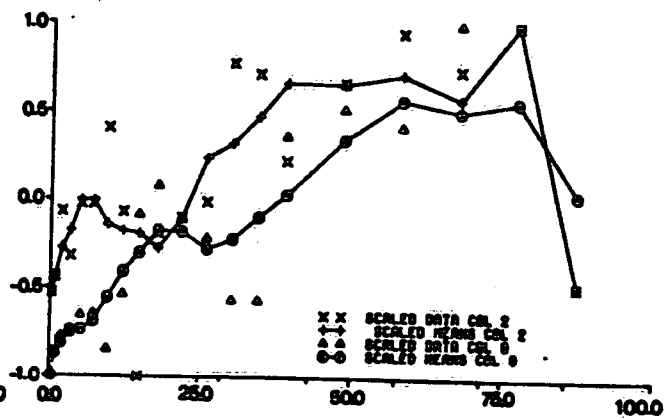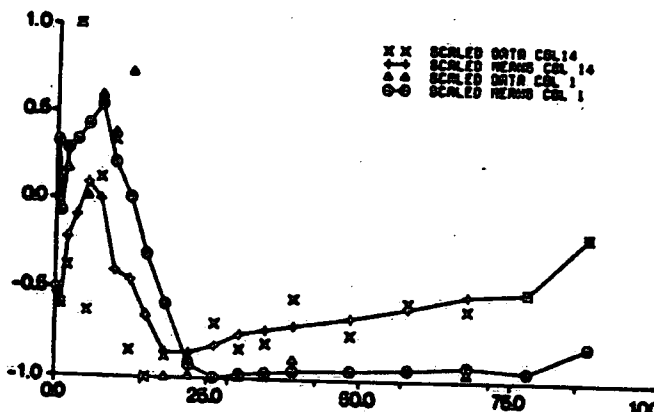
c) Data and combined spline, columns 8,2.
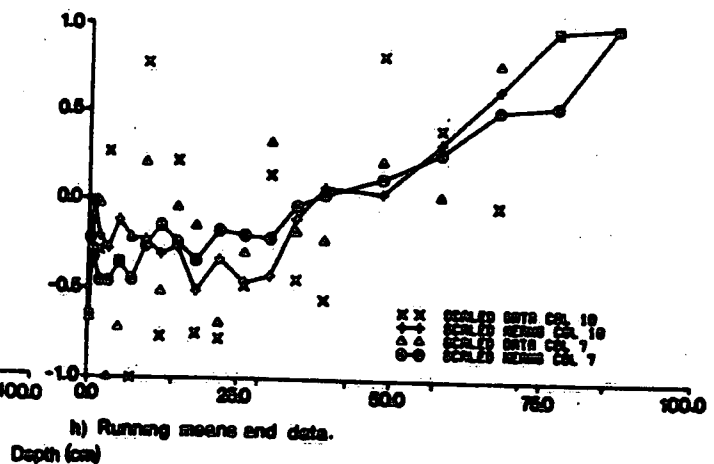
d) Data and combined spline, columns 12,4.

e) Running means and data.

f) Running means and data.

g) Running means and data.
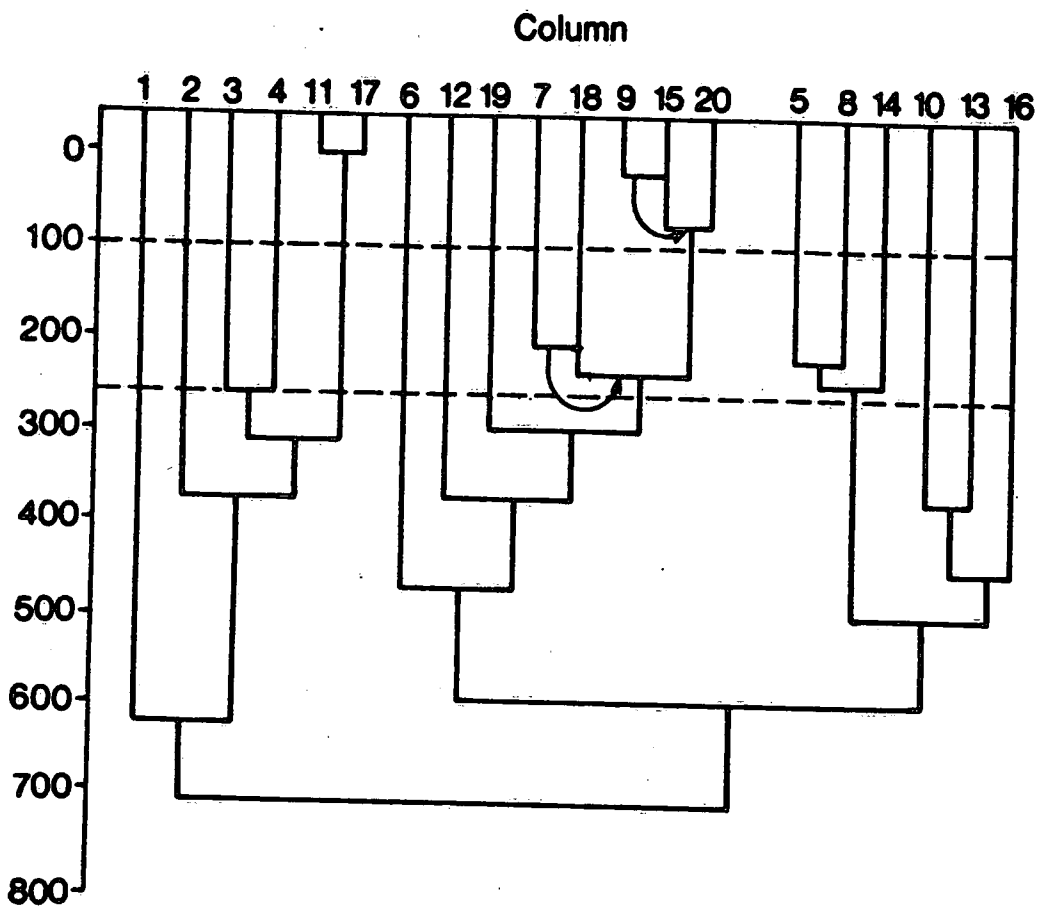
h) Running means and data.
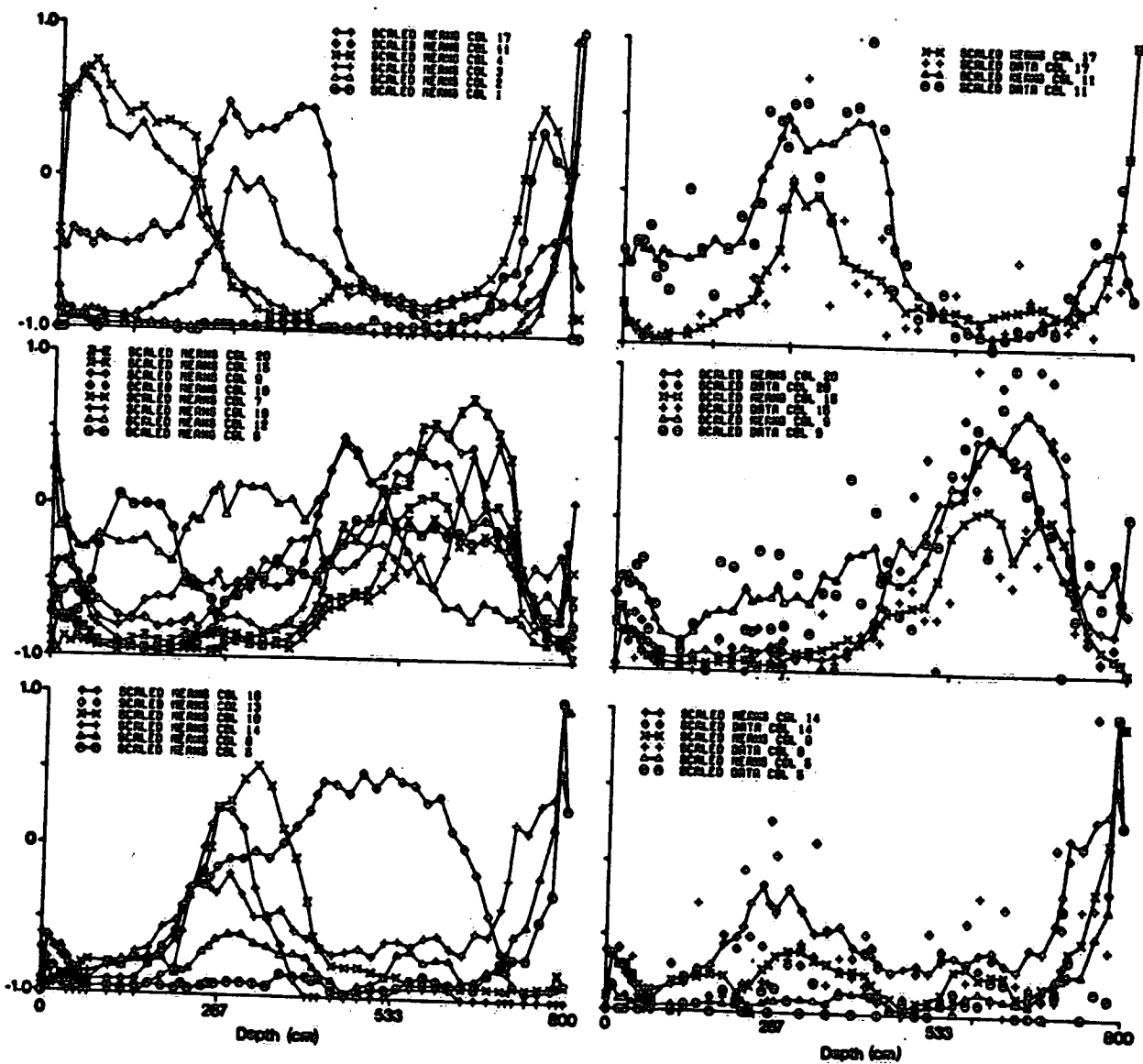
Depth (cm)

Fig. 3

Fig. 4

Fig. 5