Appl. Statist. (1987)
36, *No.* 1, *pp.* 22–33

Selection of Variables to preserve Multivariate Data Structure, using Principal Components

By W. J. KRZANOWSKI

University of Reading, UK

[Received March 1986. Revision July 1986]

SUMMARY

A common objective in exploratory multivariate analysis is to identify a subset of the variables which conveys the main features of the whole sample. Analysis of a well-known multivariate data set shows that methods currently available for selecting variables in principal component analysis may not lead to an appropriate subset. A new selection method, based on Procrustes Analysis, is proposed and shown to lead to a better subset for the data first analysed. Some supporting Monte Carlo results are presented, and implications for other multivariate techniques are briefly discussed.

Keywords: Monte Carlo simulation; Principal component analysis; Procrustes rotation; Singular value decomposition; Variable selection

1. Introduction

Exploratory multivariate studies generally aim at data inspection and dimensionality reduction. If sample features are unknown *a priori*, one of the objectives of the analysis is to uncover any interesting patterns that may be exhibited by the data. It often happens that the investigator has also measured more variables than strictly necessary on each sample member. Consequently, the dimensionality can be reduced easily without disturbing the overall features of the sample. The most common data-exploratory technique for use in such circumstances is principal component analysis [see, e.g. Mardia, Kent and Bibby (1979, Chapter 8), Seber (1984, pp. 176–203) or Jolliffe (1986)]. Suppose that p variates $X_1, X_2, \ldots, X_p$ have been observed on each of n individuals and that the observation vector for the ith individual is denoted by $X^{(i)}=(X_{i1}, X_{i2}, \ldots, X_{ip})'$. Principal component analysis linearly transforms the variates $X_1, \ldots, X_p$ to new variates $Y_1, \ldots, Y_p$, the principal components, and hence the data observations $X^{(i)}$ to corresponding principal component scores $Y^{(i)}=(Y_{i1}, \ldots, Y_{ip})'$. Dimensionality reduction is effected if $q(<p)$ of the components Y_i convey "most of the sample information" inherent in the p variates X_i. In this case the original observations $X^{(i)}$ can be replaced by the first q elements of the corresponding principal component scores. We can then write $Y^{(i)}=(Y_{i1}, \ldots, Y_{iq})'$. A plot of these scores against q orthogonal axes yields an approximation to the original data configuration. In particular, if $q=2$, the resulting scatterplot yields the best two-dimensional projection of the original high-dimensional data swarm for inspecting the sample and revealing interesting patterns.

The major deficiency of this approach to dimensionality reduction is that, while the dimensionality of the space may indeed be reduced from p to q, *all p original variables are in*

Address for correspondence: Department of Applied Statistics, The University of Reading, Box 217, Whiteknights, Reading, RG6 2AN, UK

0035–9254/87/36022 $2.00

general still needed in order to define the q new variables Y_i. This deficiency is highlighted by Srivastava and Khatri (1979) and also by McCabe (1984), who states: "In many applications, it is desirable not only to reduce the dimension of the space, but also to reduce the number of variables which are to be considered or measured in the future". One practical example of this objective occurs in sensory data analysis, where a panel of judges may be asked to rate each of a set of n food products (e.g. steak, sausage, beefburger) with respect to each of a set of p descriptors (e.g. smooth, gritty, chewy). Commonly, the objective is to quantify the perceived similarities among the products and hence to assess the market potential of a new product. When initiating such an experiment, the appropriate descriptors on which to compare the products may not be known. A possible approach is to start with a very large and exhaustive number p of such descriptors, conduct a pilot study, and use the results to select the q "best" descriptors. In such a pilot study, many of the descriptors will inevitably be redundant and the true dimensionality of the data will be much smaller than p. However, retaining all p descriptors but using only q linear combinations of them in future experiments is unsatisfactory not only because it is wasteful of time and resources but also because interpretation of such linear combinations is often difficult. Hence a reduction in the number of descriptors themselves is essential.

Identification of redundant variables, and selection of subsets of variables, is a long-standing area of interest in multivariate analysis. The most researched areas of application are those of multiple regression and discriminant analysis, where there exist natural criteria with respect to which optimality of selection can be defined (e.g. predictive sum of squares; error rate). Some references in these areas are Beale, Kendall and Mann (1967), Hocking (1976) and McKay and Campbell (1982, a,b). Only two systematic studies seem to have been made in the context of principal component analysis, however. Jolliffe (1972, 1973) discussed a number of methods for selecting a subset of q variables which preserve most of the variation of the original variables. McCabe (1984) worked with the concept of a "principal variable", which is a variable containing (in some sense) as much sample information as possible, and showed how the selection of a subset of such principal variables can be made on the basis of various optimality criteria.

From a practitioner's point of view, the best subset of q variables will contain those variables which reproduce as closely as possible the general features of the complete data (as might be evinced by a principal component analysis of the original p variables). While Jolliffe's or McCabe's subsets all satisfy various optimality criteria, they do not necessarily meet this particular requirement; this is demonstrated in section 2 for a data set first analysed by Jeffers (1967). The purpose of this paper is therefore to suggest an alternative criterion and subset selection technique which is more directly aimed at meeting the requirement. Methodological details are considered in section 3, some investigations of the technique via Monte Carlo simulations are reported in section 4, and the method is briefly placed in a wider context in section 5.

2. Motivation

Jeffers (1967) has provided two detailed case studies of principal component analysis. We focus here on the second of these. The data comprise 19 variables measured on each of 40 winged aphids (*alate adelges*) that had been caught in a light trap. The full 40 × 19 data matrix is given in Table 1. A description of the variables is given by Jeffers, and from their disparate nature it is evident that the data should be standardised before analysis. Full details of the analysis are given in Jeffers (1967); his two main conclusions were that the essential dimensionality of the data was two and that the 40 aphids formed four distinct groups. The first conclusion was based on inspection of the eigenvalues of the correlation matrix (his table 10), while the second came from the plot of the data on the first two principal components as axes (his Fig. 1).

TABLE 1

Alate Adelges data. [For description of variables, see Jeffers (1967)]

Aphid	*Variable*																		
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	21.2	11.0	7.5	4.8	5	2.0	2.0	2.8	2.8	3.3	3	4.4	4.5	3.6	7.0	4.0	8	0	3
2	20.2	10.0	7.5	5.0	5	2.3	2.1	3.0	3.0	3.2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
3	20.2	10.0	7.0	4.6	5	1.9	2.1	3.0	2.5	3.3	1	4.2	4.4	3.3	7.0	4.0	6	0	3
4	22.5	8.8	7.4	4.7	5	2.4	2.1	3.0	2.7	3.5	5	4.2	4.4	3.6	6.8	4.1	6	0	3
5	20.6	11.0	8.0	4.8	5	2.4	2.0	2.9	2.7	3.0	4	4.2	4.7	3.5	6.7	4.0	6	0	3
6	19.1	9.2	7.0	4.5	5	1.8	1.9	2.8	3.0	3.2	5	4.1	4.3	3.3	5.7	3.8	8	0	3.5
7	20.8	11.4	7.7	4.9	5	2.5	2.1	3.1	3.1	3.2	4	4.2	4.7	3.6	6.6	4.0	8	0	3
8	15.5	8.2	6.3	4.9	5	2.0	2.0	2.9	2.4	3.0	3	3.7	3.8	2.9	6.7	3.5	6	0	3.5
9	16.7	8.8	6.4	4.5	5	2.1	1.9	2.8	2.7	3.1	3	3.7	3.8	2.8	6.1	3.7	8	0	3
10	19.7	9.9	8.2	4.7	5	2.2	2.0	3.0	3.0	3.1	0	4.1	4.3	3.3	6.0	3.8	8	0	3
11	10.6	5.2	3.9	2.3	4	1.2	1.0	2.0	2.0	2.2	6	2.5	2.5	2.0	4.5	2.7	4	1	2
12	9.2	4.5	3.7	2.2	4	1.3	1.2	2.0	1.6	2.1	5	2.4	2.3	1.8	4.1	2.4	4	1	2
13	9.6	4.5	3.6	2.3	4	1.3	1.0	1.9	1.7	2.2	4	2.4	2.3	1.7	4.0	2.3	4	1	2
14	8.5	4.0	3.8	2.2	4	1.3	1.1	1.9	2.0	2.1	5	2.4	2.4	1.9	4.4	2.3	4	1	2
15	11.0	4.7	4.2	2.3	4	1.2	1.0	1.9	2.0	2.2	4	2.5	2.5	2.0	4.5	2.6	4	1	2
16	18.1	8.2	5.9	3.5	5	1.9	1.9	1.9	2.7	2.8	4	3.5	3.8	2.9	6.0	4.5	9	1	2
17	17.6	8.3	6.0	3.8	5	2.0	1.9	2.0	2.2	2.9	3	3.5	3.6	2.8	5.7	4.3	10	1	2

18	19.2	6.6	6.2	3.4	5	2.0	1.8	2.2	2.3	2.8	4	3.5	3.4	2.5	5.3	3.8	10	1	2
19	15.4	7.6	7.1	3.4	5	2.0	1.9	2.5	2.5	2.9	4	3.3	3.6	2.7	6.0	4.2	8	1	3
20	15.1	7.3	6.2	3.8	5	2.0	1.8	2.1	2.4	2.5	4	3.7	3.7	2.8	6.4	4.3	10	1	2.5
21	16.1	7.9	5.8	3.7	5	2.1	1.9	2.3	2.6	2.9	5	3.6	3.6	2.7	6.0	4.5	10	1	2
22	19.1	8.8	6.4	3.9	5	2.2	2.0	2.3	2.4	2.9	4	3.8	4.0	3.0	6.5	4.5	10	1	2.5
23	15.3	6.4	5.3	3.3	5	1.7	1.6	2.0	2.2	2.5	5	3.4	3.4	2.6	5.4	4.0	10	1	2
24	14.8	8.1	6.2	3.7	5	2.2	2.0	2.2	2.4	3.2	5	3.5	3.7	2.7	6.0	4.1	10	1	2
25	16.2	7.7	6.9	3.7	5	2.0	1.8	2.3	2.4	2.8	4	3.8	3.7	2.7	5.7	4.2	10	1	2.5
26	13.4	6.9	5.7	3.4	5	2.0	1.8	2.8	2.0	2.6	4	3.6	3.6	2.6	5.5	3.9	10	1	2
27	12.9	5.8	4.8	2.6	5	1.6	1.5	1.9	2.1	2.6	5	2.8	3.0	2.2	5.1	3.6	9	1	3
28	12.0	6.5	5.3	3.2	5	1.9	1.9	2.3	2.5	3.0	5	3.3	3.5	2.6	5.4	4.3	8	1	2
29	14.1	7.0	5.5	3.6	5	2.2	2.0	2.3	2.5	3.1	5	3.6	3.7	2.8	5.8	4.1	10	1	2
30	16.7	7.2	5.7	3.5	5	1.9	1.9	2.5	2.3	2.8	5	3.4	3.6	2.7	6.0	4.0	10	1	2.5
31	14.1	5.4	5.0	3.0	5	1.7	1.6	1.8	2.5	2.4	5	2.7	2.9	2.2	5.3	3.6	8	1	2
32	10.0	6.0	4.2	2.5	5	1.6	1.4	1.4	2.0	2.7	6	2.8	2.5	1.8	4.8	3.4	8	1	2
33	11.4	4.5	4.4	2.7	5	1.8	1.5	1.9	1.7	2.5	5	2.7	2.5	1.9	4.7	3.7	8	1	2
34	12.5	5.5	4.7	2.3	5	1.8	1.4	1.8	2.2	2.4	4	2.8	2.6	2.0	5.1	3.7	8	0	2
35	13.0	5.3	4.7	2.3	5	1.6	1.4	1.8	1.8	2.5	4	2.7	2.7	2.1	5.0	3.6	8	1	2
36	12.4	5.2	4.4	2.6	5	1.6	1.4	1.8	2.2	2.2	5	2.7	2.5	2.0	5.0	3.2	6	1	2
37	12.0	5.4	4.9	3.0	5	1.7	1.5	1.7	1.9	2.4	5	2.7	2.7	2.0	4.2	3.7	6	1	2
38	10.7	5.6	4.5	2.8	5	1.8	1.4	1.8	2.2	2.4	4	2.7	2.6	2.0	5.0	3.5	8	1	2
39	11.7	5.5	4.3	2.6	5	1.7	1.5	1.8	1.9	2.4	5	2.6	2.5	1.9	4.6	3.4	8	1	2
40	12.8	5.7	4.8	2.8	5	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5.0	3.1	8	1	2

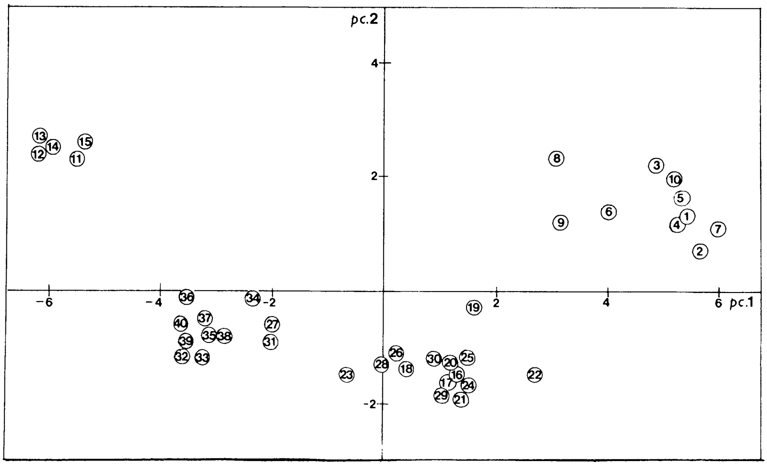

Fig. 1. *Alate* data exhibited against the first two principal components computed from all nineteen variables.

This analysis suggests that considerably fewer than 19 variables should be sufficient for identification of the structure in the data, but how many variables are necessary? Subjecting the standardised data to the cross-validatory technique described by Eastment and Krzanowski (1982) showed that four components were necessary to model the "signal" in the data, the remaining fifteen dimensions being a reflection of the "noise". This suggests that the minimum number of variables necessary for recovery of the data structure is four. Application of each of the criteria in Jolliffe (1972) and McCabe (1984) to this data set gives the following possible 4-variable subset selections for consideration:

Jolliffe	*McCabe*
(i) variables 5, 8, 11, 14	(iii) variables 9, 11, 17, 19
(ii) variables 5, 11, 13, 17	(iv) variables 5, 9, 11, 18
	(v) variables 6, 11, 17, 19
	(vi) variables 5, 8, 11, 18

Some further details about these subsets may be found in Jolliffe (1986, p. 111). For present purposes, we wish to determine how well each subset represents the structure of the complete data. To do this, a principal component analysis was first done on the complete 40 × 19 (standardised) data matrix, and the (40 × 4) matrix of scores on the first four principal components was obtained. The first four principal components account for 91.9 % of the total variation, so this matrix is a good approximation to the overall data matrix. The (40 × 4) data matrix obtained from each of the above subsets was then compared in turn with this (40 × 4) matrix of scores, by means of Procrustes Analysis (see next section). The residual sums of squares for the six analyses provide a ranking of the six subsets; the smaller the sum of squares, the better the agreement between the subset data and the overall data. The six residual sums of squares were (i) 239, (ii) 258, (iii) 267, (iv) 258, (v) 266 and (vi) 261; the best Jolliffe subset comprised variables 5, 8, 11, 14, while the best McCabe subset comprised variables 5, 9, 11, 18. To determine how well these two subsets capture the structure of the complete data, a further principal component analysis was done on each (40 × 4) data matrix and the data were plotted on the first two principal components as axes. As a benchmark, Fig. 1 shows the two-dimensional representation of the complete data matrix while Fig. 2 presents the two-dimensional representation for the Jolliffe subset 5, 8, 11, 14 and Fig. 3 the two-dimensional representation for the McCabe subset 5, 9, 11, 18. Each two-dimensional

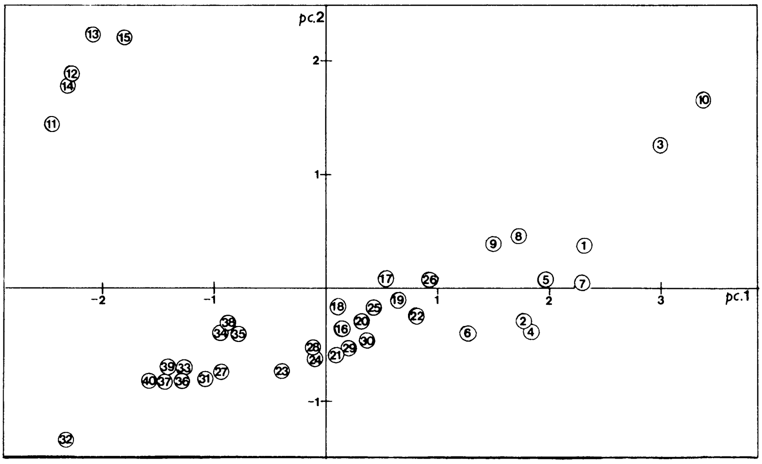

Fig. 2. *Alate* data exhibited against the first two principal components computed from variables 5, 8, 11 and 14 only.

representation is a good approximation to the relevant complete matrix, with 85.3, 82.3 and 78.2 per cent variance accounted for in Figs 1, 2 and 3 respectively. Comparison of Figs 2 and 3 with Fig. 1 shows that, while the tight group of aphids in the top left corner of the full analysis is readily identified (albeit more loosely) in the two subset analyses, the three groups in the lower part of the full analysis have been merged. In Fig. 3 these groups are indistinct, while Fig. 2 is between Figs 1 and 3 in terms of data structure. None of the other subsets provides any improvement in this respect, and hence none of the subsets adequately captures the group structure of the complete data.

Thus it is evident that a different criterion to those previously defined is necessary, if capturing the data structure is the prime objective. This criterion is considered in the next section.

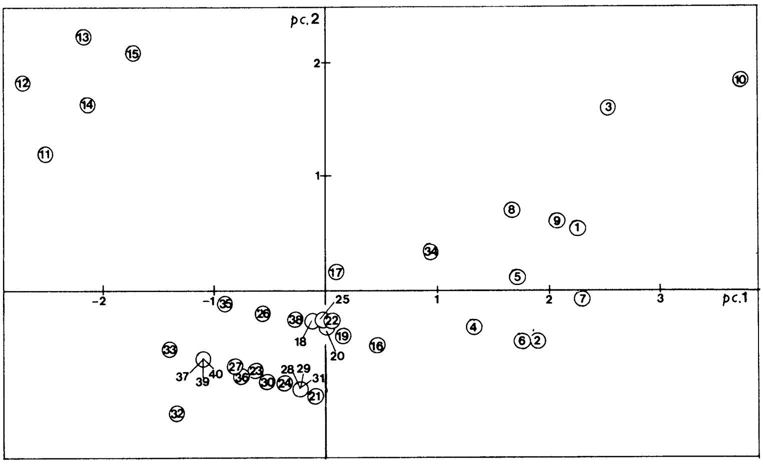

Fig. 3. *Alate* data exhibited against the first two principal components computed from variables 5, 9, 11 and 18 only.

3. A Procrustes Criterion

One possible reason why none of the selection methods considered in section 2 identified a subset of variables that recovered the group structure of the complete data is that all the selection criteria are concerned with overall features, either of the subset data (McCabe) or of the complete data (Jolliffe). Thus the criteria are based exclusively on variance-covariance or correlation matrices and their eigenvalues/eigenvectors. A more appropriate criterion for preserving structure among units will be one that involves some direct comparisons between the individual points of the subset configuration and the corresponding points of the complete data configuration. Since the same number of units are involved in both the complete data and the subset data, a criterion based on Procrustes Analysis (Gower, 1971; Sibson, 1978) seems to be a natural contender. Derivation of such a criterion will now be outlined.

Denote by X the $(n \times p)$ data matrix (standardised if appropriate), consisting of the values of p variables measured on each of n sample individuals. Suppose that the essential dimensionality of the data, to be used in any comparison, is k. This may be decided either formally [e.g. by using the cross-validatory techniques of Wold (1978) or Eastment and Krzanowski (1982)] or informally from considerations of convenience (e.g. because two dimensions have been decided upon for displaying the data). However, in the latter case, care must be taken to ensure that sufficient data variability has been accounted for in the chosen k. Let Y be the $(n \times k)$ transformed data matrix of principal component scores, yielding the best k-dimensional approximation to the original data configuration X. We would now like to select q of the original p variables, where $q < p$ (so that selection does take place) and $q \geqslant k$ (so that there is hope of recovering the true structure with the selected variables). Suppose that $\tilde{X}$ denotes the $(n \times q)$ data matrix which retains only q selected variables $X_u, X_v, \ldots, X_w$, and that $\tilde{Z}$ is the $(n \times k)$ matrix of principal component scores of these reduced data. This latter is therefore the best k-dimensional approximation to the q-dimensional configuration defined by the subset data. If the true dimensionality of the data is indeed k, then Y can be viewed as the "true" configuration, while $\tilde{Z}$ is the corresponding approximate configuration based on only the q variables $X_u, X_v, \ldots, X_w$. To measure the discrepancy between the latter configuration and the true configuration we can conduct a Procrustes Analysis, viz. find the sum of squared differences between corresponding points of the two configurations after they have been matched as well as possible under translation, rotation and reflexion. This residual sum of squares therefore measures the loss of information about the data structure when only the q variables $X_u, X_v, \ldots, X_w$ are used instead of all p variables. The schematic diagram below shows the steps of the procedure.

$$\begin{tikzcd}
X(n \times p) \arrow[r, "Select"] \arrow[d, "PCA"] & \tilde{X}(n \times q) \arrow[d, "PCA"'] \\
Y(n \times k) \arrow[r, leftrightarrow, "Procrustes"'] & \tilde{Z}(n \times k)
\end{tikzcd}$$

Matching under translation is effected by ensuring that both Y and $\tilde{Z}$ are mean-centred (Gower, 1971; Sibson, 1978). Matching under rotation and reflexion is effected by considering one configuration to be fixed and transforming the other one under these operations. Since we wish to find the loss of information caused by deletion of some variables, Y is the natural choice for the fixed configuration while $\tilde{Z}$ is the configuration that is to be transformed. Standard results of Procrustes Analysis (e.g. Gower, 1971) then yield the sum of squared differences between the two configurations as

$$M^2 = \mathrm{Trace}\{YY' + \tilde{Z}\tilde{Z}' - 2\tilde{Z}Q'Y'\}. \tag{1}$$

In this expression, Q is given by multiplying together two of the matrices in the singular value decomposition of the $k \times k$ square matrix $\tilde{Z}'Y$. This takes the form (see, e.g., Good, 1969):

$$\tilde{Z}'Y = U\Sigma V'$$

where $U'U = I_k$, $V'V = VV' = I_k$ and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$. Computer algorithms for calculating U, Σ and V are readily available (e.g. Golub and Reinsch, 1970). The matrix Q is then given by $Q = VU'$. Substituting into (1) and carrying out some algebraic manipulation yields

$$M^2 = \mathrm{Trace}\{YY' + \tilde{Z}\tilde{Z}' - 2\Sigma\} \qquad (2)$$

The quantity M^2 given by equation (2) can be computed readily for any selected subset X_u, $X_v, \ldots, X_w$ of the variables, and denotes the closeness with which the configuration based on these variables matches the configuration derived from the complete data. The "best" subset of q variables will thus be that subset which yields the smallest value of M^2 among all q-variable subsets. In principle, subset selection can be effected by computing (2) for all possible subsets of variables and choosing the ones which give minimum M^2 for each desired choice of q. In practice, of course, this procedure may not be computationally feasible, and an alternative strategy needs to be employed. Any of the standard approaches (forward selection, backward elimination or stepwise selection) can be implemented, but forward and stepwise selection lead to time consuming procedures as well as posing various programming problems. On the other hand, given some very efficient algorithms for obtaining an amended singular value decomposition after the deletion of a single variable from a data set (Bunch and Nielsen, 1978; Bunch, Nielsen and Sorensen, 1978), the following backward elimination procedure can be programmed easily and has proved to be very effective.

(I) Initially set $q = p$, and for fixed k compute the matrix of principal component scores Y. Set $Z = Y$.
(II) Using the updating algorithms in Bunch *et al*, obtain and store the matrix of principal component scores on deleting in turn each variable from Z.
(III) Compute M^2 from equation (2) for each such matrix of scores and identify the variable X_u which yields the smallest M^2. Let $\tilde{Z}_{(u)}$ denote the corresponding matrix of scores.
(IV) Delete variable X_u. Set $Z = \tilde{Z}_{(u)}$ and return to stage II with p–1 variables. Continue this cycle until only q variables are left.

This process is computationally very fast. Step I requires one complete singular value decomposition, to obtain the principal component scores Y. (If the singular value decomposition of the original data matrix X is given by $X = U\Sigma V'$, where U is $n \times p$ while Σ and V are both $p \times p$, then Y is given by the first k columns of $U\Sigma$.) Step II requires the fast updating subroutine for each variable deletion. (If $\bar{U}$ and $\bar{\Sigma}$ are the amended elements of the earlier singular value decomposition on deletion of a given variable, then the amended matrix of scores is given by the first k columns of $\bar{U}\bar{\Sigma}$.) Step III requires only k singular values (elements of Σ) to be computed for each application of equation (2). No singular vectors (columns of U or V) are required in this step, which improves considerably the speed of computation.

Returning to the *alate* data, the above backward elimination procedure yielded the subset containing variables 5, 12, 14 and 18 (using the value $k = 4$ obtained from prior application of the Eastment/Krzanowski technique). The value of M^2 for this subset was 221, which is substantially lower than any of the values for the previous subsets (i)–(vi). Principal component analysis of this subset gave the two-dimensional representation shown in Fig. 4 (with 90.4% of variability accounted for by the first two components). It can be seen that the four groups of the complete data are recovered very well. In fact the variable selection picks out those variables most important for defining the groups. In the absence of the fifteen other variables that create the "noise" in the system the groups in the subset data are even more sharply delineated than they are in the full data. There might even be a case for considering the

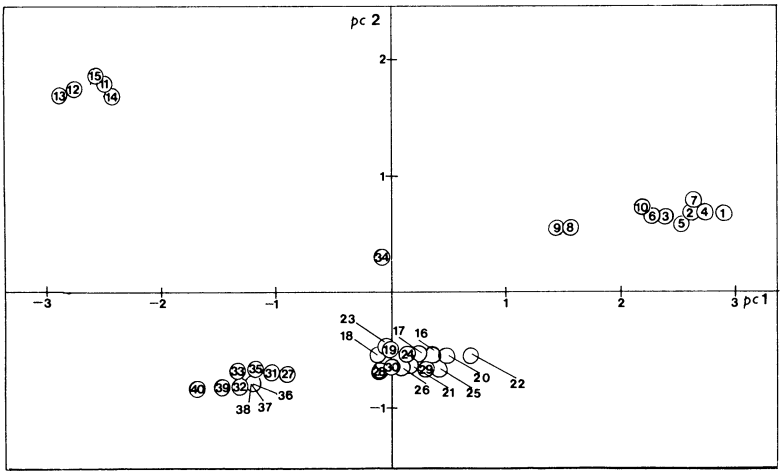

Fig. 4. *Alate* data exhibited against the first two principal components computed from variables 5, 12, 14 and 18 only.

existence of two extra "groups", consisting of aphid numbers 34 and 8 + 9 respectively. Aphid number 34 has untypical value 0 on variable 18 whereas all other aphids in its group of Fig. 1 have value 1. Aphids 8 and 9 have distinctly lower values for variables 12, 13 and 14 than the others in their group of Fig. 1.

4. Some Monte Carlo Results

The analysis of the *alate* data has demonstrated the value of the Procrustes criterion, and analyses of various sets of sensory data have also led to sensible choices of subsets. Some of these analyses will be reported elsewhere. To investigate the whole procedure more broadly, however, a Monte Carlo simulation study was set up along the lines of the one described by Jolliffe (1972). Jolliffe generated a large number of artificial data sets, conforming to one of five predetermined models. Each model was constructed in such a way that certain variables were linear combinations of other variables, except for a random disturbance, and hence were redundant. He then tested his various rejection methods on the data to see whether the variables they rejected were the redundant ones. In all his models, there was some choice regarding the best set of rejected variables, and he accordingly labelled the different possible subset selections as "best", "good", "moderate" or "bad". He gave in his table 2 a definition of the constructed variables for each of models I–IV, and specified this subset labelling for each of the same models in his table 3. Model V was more complicated, and was described separately.

In the present study, data were generated in accordance with each of Jolliffe's models I–IV. One hundred samples of size 100 were generated for each of these models, and the procedure described in Section 3 selected the following subsets.

Model I	Subset (4, 5, 6)	100 times ("Best" in Jolliffe's classification)
Model II	Subset (3, 4, 6)	70 times ("Good")
	(3, 4, 5)	11 times ("Good")
	(1, 3, 6)	1 time ("Good")
	(4, 5, 6)	18 times ("Bad")
Model III	Subset (4, 5, 6)	100 times ("Good")
Model IV	Subset (1, 3, 6, 10)	91 times ("Best")
	(3, 6, 9, 10)	8 times ("Bad")
	(3, 5, 6, 10)	1 time ("Bad")

In each simulation, the "true" dimensionality for comparison (k of Section 3) was set to be the true dimensionality of the model, i.e. 3 for models I, II and III and 4 for model IV, and a subset of k variables was selected each time.

The main drawback with this simulation study is that the data sets are purely random, each of the generated variables either being a $N(0, 1)$ variable or a linear combination of $N(0, 1)$ variables. There is therefore no inherent structure present, each data set being a collection of points randomly scattered about zero. Jolliffe's main objective was to investigate the rate at which the various methods succeeded in discarding the redundant variables. From the results above it can be seen that the method based on the Procrustes criterion performs creditably in this respect. The poorest performance is for Model II. This is the only model which contains one variable that is independent of the rest, counterbalanced by a group of three highly intercorrelated variables. It can be noted that neither of Jolliffe's two rejection methods selected any "best" subsets for this model, either. However, his inclusion method did select the "best" subset nearly always. On the other hand, the Procrustes criterion clearly outperforms Jolliffe's inclusion method for models I and III, is only marginally worse than the inclusion method for model IV, and is only clearly outperformed by either of Jolliffe's rejection methods for model II. However, these data sets do not test whether or not the selected variables preserve the essential features of the data, and indeed Jolliffe's subset classifications do not reflect this objective. To do this, it is necessary to build in some extra structure into the data, and then see whether the variables that are selected are the ones that carry this structure.

Extra structure was therefore built into Jolliffe's models, to create clear groupings of individuals. This was done in a 3^2 factorial manner as follows. The first factor controlled the *type* of structure present, with levels (single outlier, weak groups, strong groups), while the second factor controlled the *amount* of structure present with levels (in one variable, in two variables, in three variables). In the case of the "single outlier", the value 10 was added to each of the first j variables of the first sample member, where j was the chosen level of the second factor. Thus the first sample individual differed from the remaining sample members in mean value by ten sample standard deviations on each of j variables. In the case of "weak groups", the first 25 sample members were left unchanged, the next 25 sample members had the value 2 added to each of the first j variables, the following 25 sample members had the value 4 added to each of the first j variables, while the last 25 sample members had the value 6 added to each of the first j variables. For the "strong groups" level the same procedure was carried out except that multiples of 10 replaced multiples of 2 when the constant value was added. Once again, j was the chosen level of the second factor in each case. Variable selection was then carried out using the procedure described in Section 3, k again being set to the true dimensionality of the chosen model. The composition of the chosen subset, as regards the number of structure-carrying variables that had been correctly selected, was recorded and the results are given in Table 2. If $j = 1$ for the second factor, then the tabulated results give the number of times that variable 1 was selected as part of the chosen subset for each model, as well as the number of times that variable 1 did not appear in the chosen subset. If $j = 2$ then the tabulated results comprise the number of times both variables 1 and 2 appeared in the chosen subset, the number of times either variable1 or variable 2 appeared and the number of times that neither appeared. Finally, if $j = 3$, the tabulated results comprise the number of times all three variables 1, 2 and 3 appeared in the chosen subset, the number of times two of the three appeared, the number of times one of the three appeared, and the number of times none appeared.

As the structure in the data becomes progressively stronger, so does it become more likely that the structure-bearing variables are included in the chosen subset. With strong grouping, all chosen subsets contained all of these variables. Note in particular that to include all of variables 1, 2 and 3 in the case $j = 3$ for model IV, the in-built redundancy between variables 2 and 3 has first to be overcome. With weak grouping the structure is not strong enough to

TABLE 2

Results of Monte Carlo variable selections. Number of times the given number of structure variables is selected for each model and 3^2 factorial design. For definition of models see Jolliffe (1972)

Structure	*Number of variables exhibiting structure*	*Number of structure variables selected*	*Model I*	*Model II*	*Model III*	*Model IV*
Single Outlier	1	0	0	0	9	0
		1	100	100	91	100
	2	0	0	0	2	0
		1	6	15	75	97
		2	94	85	23	3
	3	0	0	0	0	0
		1	6	20	6	0
		2	84	72	84	100
		3	10	8	10	0
Weak Grouping	1	0	0	0	0	0
		1	100	100	100	100
	2	0	0	0	0	0
		1	0	0	0	0
		2	100	100	100	100
	3	0	0	0	0	0
		1	0	0	0	0
		2	1	0	11	100
		3	99	100	89	0
Strong Grouping	1	0	0	0	0	0
		1	100	100	100	100
	2	0	0	0	0	0
		1	0	0	0	0
		2	100	100	100	100
	3	0	0	0	0	0
		1	0	0	0	0
		2	0	0	0	0
		3	100	100	100	100

do this, and for $j = 3$ all chosen subsets contained only two of variables 1, 2, 3 under model IV. Models I and III also contained a small proportion of data sets in which this happens. Apart from these instances, however, all chosen subsets contained the correct number of structure-bearing variables. Hence, under group structure, the procedure of Section 3 is expected to perform well. When the only "structure" in the data comes from a single outlier, however, the results are more variable. The most notable pattern that emerges is that one variable fewer than the number defining the outlier is identified. Thus when $j = 3$, two variables are selected consistently for all models; when $j = 2$ both variables are well identified in models I and II, but only one is consistently chosen in the other two moels. When $j = 1$, however, then variable 1 is chosen almost always by all models.

5. Comment

The Monte Carlo simulations of the previous section, and the real data sets analysed to date, show that the technique described in this paper will identify structure-bearing variables in a data set and successfully isolate them from "noise" variables, particularly when groups are present in the data. This has implications for a number of other multivariate techniques such as discriminant analysis, projection pursuit and cluster analysis. First, however, we must carefully distinguish our objectives from those of the latter techniques. We are attempting to select a subset of variables that preserve what (*unknown*) structure may be present in the data. If we wish to *use* a known group structure in further analysis, then discriminant analysis is appropriate; if we wish to *find* groups or outliers in the data, then either cluster analysis or projection pursuit (Huber, 1985) is appropriate. Selection of variables sometimes features as an adjunct to each of these techniques, and the method of this paper may be contrasted with such additional selection exercises. In discriminant analysis, attention is usually focussed on the selection of variables in order to maximise the probability of correct allocation to existing groups of future individuals (a *classification* objective), while the present technique provides a method for selecting those variables which best show up the presence of the existing groups (a *discrimination* objective). The former objective usually requires some probabilistic assumptions, while the latter is strictly data-based. In cluster analysis or projection pursuit, once the sample has been partitioned into clusters, the practitioner is often interested in identifying a small number of variables which can be used to describe cluster membership and distinguish between clusters. If the data do form groups, then the technique of this paper will accomplish this objective directly. Indeed, the variables selected for the *alate* data (Figure 4) almost provide a key to classification in that each variable distinguishes the individuals in one of the four groups from the rest of the sample almost perfectly. The present technique may thus also provide a useful tool in setting up diagnostic keys (see Payne and Preece, 1980).

Acknowledgements

I thank Mr. J. N. R. Jeffers and Dr. C. Carter for access to the full *alate* data set and for agreeing to its publication. I am also grateful to Dr. B. J. T. Morgan, Dr. I. T. Jolliffe, and two referees, for a number of comments and suggestions which helped to improve the presentation.

References

Beale, E. M. L., Kendall, M. G. and Mann, D. W. (1967) The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357–366.

Bunch, J. R. and Nielsen, C. P. (1978) Updating the singular value decomposition. *Numerische Mathematik*, **31**, 111–129.

Bunch, J. R., Nielsen, C. P. and Sorensen, D. C. (1978) Rank one modification of the symmetric eigenproblem. *Numerische Mathematik*, **31**, 31–48.

Eastment, H. T. and Krzanowski, W. J. (1982) Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, **24**, 73–77.

Golub, G. H. and Reinsch, C. (1970) Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**, 403–420.

Good, I. J. (1969) Some applications of the singular value decomposition of a matrix. *Technometrics*, **11**, 823–831.

Gower, J. C. (1971) Statistical methods of comparing different multivariate analyses of the same data. In *Mathematics in the Archaeological and Historical Sciences* (F. R. Hodson, D. G. Kendall and P. Tautu, eds.) pp. 138–149. Edinburgh: University Press.

Hocking, R. R. (1976) The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–50.

Huber, P. J. (1985) Projection pursuit. *Ann. Statist.*, **13**, 435–475.

Jeffers, J. N. R. (1967) Two case studies in the application of principal component analysis. *Appl. Statist.*, **16**, 225–236.

Jolliffe, I. T. (1972) Discarding variables in a principal component analysis. I: Artificial data. *Appl. Statist.*, **21**, 160–173.

Jolliffe, I. T. (1973) Discarding variables in a principal component analysis. II: Real data. *Appl. Statist.*, **22**, 21–31.

Jolliffe, I. T. (1986) *Principal Component Analysis*. New York: Springer-Verlag.

McCabe, G. P. (1984) Principal Variables. *Technometrics*, **26**, 137–144.

McKay, R. J. and Campbell, N. A. (1982a) Variable selection techniques in discriminant analysis. I. Description. *Brit. J. Math. Statist. Psychol.*, **35**, 1–29.

McKay, R. J. and Campbell, N. A. (1982b) Variable selection techniques in discriminant analysis. II. Allocation. *Brit. J. Math. Statist. Psychol.*, **35**, 30–41.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.

Payne, R. W. and Preece, D. A. (1980) Identification keys and diagnostic tables: a review (with discussion). *J. R. Statist. Soc. A*, **143**, 253–292.

Seber, G. A. F. (1984) *Multivariate Observations*. New York: Wiley.

Sibson, R. (1978) Studies in the robustness of multidimensional scaling. *J. R. Statist. Soc.* B, **40**, 234–238.

Srivastava, M. S. and Khatri, C. G. (1979) *An Introduction to Multivariate Statistics*. New York: North-Holland.

Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics*, **20**, 397–405.