Discarding Variables in a Principal Component Analysis. II: Real Data

By I. T. Jolliffe

University of Kent at Canterbury

Summary

In this paper it is shown for four sets of real data, all published examples of principal component analysis, that the number of variables used can be greatly reduced with little effect on the results obtained. Five methods for discarding variables, which have previously been successfully tested on artificial data (Jolliffe, 1972), are used. The methods are compared and all are shown to be satisfactory for real, as well as artificial, data, although none is shown to be overwhelmingly superior to the others.

Keywords: CLUSTER ANALYSIS; DISCARDING VARIABLES; MULTIPLE CORRELATION; PRINCIPAL COMPONENT ANALYSIS

1. Introduction

In a previous paper (Jolliffe, 1972) the author discussed several proposed methods for discarding variables in a principal component analysis. Artificial data were generated containing variables known to be redundant, and it was shown that for most of the data several rejection methods discarded precisely the redundant variables. In the present paper the successful methods for the artificial data are tested on four sets of real data, all published examples of component analysis.

Of the four examples two are given by Jeffers (1967), one by Ahamad (1967) and one by Moser and Scott (1961). These four examples use component analysis for a variety of purposes, so much of the comparison between results for the full and reduced sets of variables differs for each example. The results are discussed separately for each of the four sets of data in Sections 4–7 of the paper.

There is, however, one form of comparison which is used for all four examples. This is a straightforward comparison between the principal components themselves for the full and reduced sets of data. Two numerical measures used to make these comparisons are introduced in Section 3.

Five methods are used for discarding variables, and of these one is a multiple correlation method, two are principal component methods and two are clustering methods. These methods are described briefly in the next section.

2. Rejection Methods

The five rejection methods used in this paper will be labelled in the same way as in Jolliffe (1972), where a more detailed description of them may be found. The methods are of three types, as follows:

(a) The multiple correlation method, referred to as A2. This is a step-wise method, which rejects at each stage the variable having the largest multiple correlation with the remaining variables. The process stops when all multiple correlations first fall below 0·15. This value, like the others below, was chosen empirically from artificial data.

(b) The principal component methods, known as B2 and B4. These methods associate a variable with each of the principal components (in a different way for each method) and reject those variables associated with the last few components. The number of variables rejected equals the number of eigenvalues, associated with the components, which are smaller than 0·70.

(c) The clustering methods, C1 and C2. These methods place the variables in groups and retain precisely one variable from each group. One method, C1, is a single-linkage method and the other, C2, is an average-linkage method. Clustering is agglomerative and stops when all between-cluster similarities first fall below 0·45 for C2 and 0·55 for C1.

Both clustering methods may be subdivided according to the way in which one variable is selected from each group. Two methods of selection are considered here, and are referred to as outer- and inner-clustering. In outer-clustering the last variable to join a group is selected but in inner-clustering the selected variable is one of the original members of the group.

3. Comparison of Principal Components

Two types of measure are used to compare principal components for the full and reduced sets of data. One is based on product-moment correlation coefficients and the other on rank correlation coefficients.

Suppose the full set of data contains k variables, $x_1, x_2, ..., x_k$ measured on n individuals. In all four sets of data discussed below, the component analysis is done on the correlation matrix rather than the dispersion matrix, so it can be assumed, without loss of generality, that the variables each have sample mean zero, and sample variance one. Also the sample correlations r_{ij} between each pair of variables $\{x_i, x_j\}$ are known.

Any principal component for a full set of data is a linear combination of all the variables and can be written as

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_k x_k,$$

where the a_i's are constants. Similarly, any principal component for a reduced set of data can be written

$$z = b_1 x_1 + b_2 x_2 + \ldots + b_k x_k,$$

where the b_i's are constants, but here all those b_i's corresponding to rejected variables are zero.

With n values for each of the x's, n corresponding values can be computed for y and z, and the correlation coefficient between y and z calculated. This correlation coefficient, denoted r_1, can be easily shown to have the form

$$r_1 = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j r_{ij}\right) \Big/ \left\{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j r_{ij}\right)\left(\sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j r_{ij}\right)\right\}^{\frac{1}{2}}.$$

An alternative to r_1 as a measure of similarity between components, which is easier to calculate for large sets of data, is a rank correlation coefficient, r_2. If $\{C_{yi}\}$, $\{C_{zi}\}$ are the ranks of the n individuals with respect to y and z, then r_2 may be defined using the well-known formula

$$r_2 = 1 - 6\left\{\sum_{i=1}^{n} (C_{yi} - C_{zi})^2\right\} \Big/ \left\{n(n^2 - 1)\right\}.$$

Both r_1 and r_2 are used to measure similarity between a particular pair of components; to measure similarity between two sets of components a weighted average of r_1's and r_2's for several pairs of components may be used.

Usually only the first few, say m, principal components are of interest for the full set of data. Define Q_1 and Q_2, two measures of similarity between components for the full and reduced sets of data by

$$Q_j = \left(\sum_{i=1}^{m} q_i r_j(i)\right) \bigg/ \left(\sum_{i=1}^{m} q_i\right), \quad j = 1, 2,$$

where $r_j(i)$ is the maximum value of r_j, defined above, between the ith component for the full set of data and any component for the reduced set, and q_i is the proportion of the total variation accounted for by the ith component in the full set of data.

Thus $Q_j, j = 1, 2$, is a weighted sum of r_j's and the weights are proportional to the relative importance of the first few components in the full set of data.

For the first three sets of data discussed below, r_1 and Q_1 are used as similarity measures, but for the fourth set, which is much larger, it was found more convenient to use r_2 and Q_2.

4. Physical Properties of Pitprops

These data consist of 13 variables measured on 80 pitprops of Corsican pine. A full list of the variables, and also the correlation matrix, may be found in Jeffers (1967).

Each of the five rejection methods suggest that six of the original 13 variables should be retained, and the variables selected by each method are given in Table 1.

Table 1

Variables retained by rejection methods for pitprops data

Rejection method	*Variables retained*								
A2			3	5		9	11	12	13
B2	1		3	5	8		11		13
B4		2	3	5			11	12	13
C1 and C2 (inner-clustering)	1		3	5			11	12	13
C1 and C2 (outer-clustering)			3	5	8		11	12	13

Although the sets of variables retained differ for the various methods there are strong similarities and variables 3, 5, 11 and 13 are retained by all the methods.

In Jeffers (1967) it is also suggested that only six variables will be needed in similar future work, and the variables specified there are the same as given by the inner-clustering methods. However, I understand that the choice was, in fact, made using a principal component method together with experience of other similar forestry data.

The first step in comparing results for the full and reduced sets of data is to measure the similarity between the principal components for 13 variables and those for the various sets of six variables. The measures of similarity, $r_1(j), j = 1, \ldots, 6$, and Q_1 based on the first six components, are given in Table 2 for each reduced set of variables.

Also given in Table 2 is the number of the component from the reduced set of data most highly correlated with the jth component from the full set of data. This number is informative since it is often different from j for two reasons. First, the components for the full and reduced sets, although similar, may be in a different order, e.g. the set chosen by B4. Second, the components for one set of data may be "mixtures" of those for the other set, and more than one of the original components may be most highly correlated with the same "reduced component", e.g. the set chosen by A2.

TABLE 2

Measures of similarity, $r_1(j)$, Q_1, between components for the full set of pitprops data and for reduced sets of data chosen by various rejection methods

Rejection method	*A*2	*B*2	*B*4	*C*1 *and C*2 (*inner-clustering*)	*C*1 *and C*2 (*outer-clustering*)
$r_1(1)$	0·634 (2)†	0·661 (2)	0·663 (2)	0·695 (2)	0·367 (2)
$r_1(2)$	0·873 (1)	0·641 (1)	0·869 (1)	0·873 (1)	0·828 (1)
$r_1(3)$	0·697 (3)	0·606 (4)	0·679 (4)	0·715 (4)	0·751 (3)
$r_1(4)$	0·634 (2)	0·820 (3)	0·628 (3)	0·631 (3)	0·716 (2)
$r_1(5)$	0·743 (6)	0·400 (6)	0·782 (6)	0·579 (6)	0·591 (5)
$r_1(6)$	0·892 (5)	0·948 (5)	0·946 (5)	0·957 (5)	0·587 (4)
Q_1	0·720	0·663	0·734	0·736	0·591

† In brackets beside $r_1(j)$ is the number of the component from the reduced set of data most highly correlated with the jth component from the full set of data.

Thus there are three criteria by which similarity between components from the full and reduced sets of data may be judged. First, the overall values of the $r_1(j)$'s and Q_1, second the extent to which the "reduced components" are mixtures of the original components, and third the extent to which the ordering of components is changed. The last of these is least important.

Judged on values of Q_1 the set of variables chosen by the outer-clustering methods is worse than the others. This is mainly because none of the components for this reduced set has a high correlation with the first, most important, original component. The other methods all have similar values of Q_1, close to 0·7, but for method A2 the second reduced component is a mixture of two of the original components. This leaves the principal component methods and the inner-clustering methods as the most successful in reproducing the original components with a reduced set of variables, but even here the reduced components are in a different order.

A more informative comparison of rejection methods can be obtained by using the reduced components in the same way as Jeffers (1967) uses the original components, i.e. a regression analysis is done with an additional variable, maximum compressive strength, as the dependent variable and the first 6 principal components as orthogonal regressor variables.

Some of the results of this analysis, namely the percentage variation of maximum compressive strength accounted for by each component, are given in Table 3 for the full set and various reduced sets of variables.

It can be seen that, when the components are suitably re-ordered, the results for method B4 are most similar to those for the original data. The other principal component method and the inner-clustering methods also produce results close to those for the full set of data, but the multiple correlation and outer-clustering methods are not so good, even with further re-ordering of components. The relative merits of the methods are thus generally the same here as in the direct comparison of components.

TABLE 3

Percentage contribution of original and reduced principal components to variability of maximum compressive strength

		Reduced components				
Component number†	*Original components*	*A*2	*B*2	*B*4	*C*1 *and C*2 (*inner-clustering*)	*C*1 *and C*2 (*outer-clustering*)
1	33·27	25·56	43·12	33·44	37·24	21·68
2	7·56	8·95	0·16	3·85	4·84	9·78
3	0·26	0·10	0·29	0·64	2·90	0·15
4	3·05	10·95	6·10	5·42	2·70	0·00
5	6·13	12·40	9·83	3·41	2·00	26·90
6	13·70	0·17	5·82	10·50	8·47	6·96

† This ordering applies to all the reduced components; the original components are given in the order 2, 1, 4, 3, 6, 5 to make comparisons easier.

The fact that here, and in the later examples, several different subsets of variables give reasonable results illustrates a situation similar to that found in multiple regression. For any chosen criterion (e.g. R^2 in multiple regression) there is usually a unique subset of the variables, of given size, which maximizes the value of that criterion. However, there may be several other subsets of the same size for which this maximum value is nearly attained. Thus it does not matter much if one of these subsets, rather than the best, is chosen.

Finally, in this section, it is worth mentioning that Jeffers gives interpretation to each of the first six components, e.g. the first two are interpreted as measuring size of prop and degree of seasoning. Apart from the changes in the order of the components these interpretations are still valid for the components produced by the inner-clustering and principal component methods, but the interpretations are less clear for the other methods.

5. Variation of Alate Adelges

These data are also given by Jeffers (1967) and consist of 19 variables measured on 40 alate adelges (winged aphids). Jeffers does a principal component analysis on these data and plots 40 points in two dimensions, corresponding to the values, for each aphid, of the first two principal components. The object was to discover how many distinct groups, possibly species, of aphids were in the sample, and the plotted data show four major groups.

Most of the variables are closely related to the size of the aphid, and correlations between variables are large. Of the rejection methods discussed in the present paper the multiple correlation and principal component methods suggest that only three variables need be retained, whereas the clustering methods suggest only two variables are necessary. Subsets of three selected variables, together with values of $r_1(j)$, $j = 1, 2, 3$, and Q_1 based on the first three components, are given in Table 4 for the various rejection methods.

TABLE 4

Measures of similarity, $r_1(j)$, Q_1, between components for the full set of Alate adelges data and for reduced sets of three variables chosen by various rejection methods

Rejection method	*A*2	*B*2	*B*4	*C*1 (*inner*)	*C*2 (*inner*)	*C*1 *and C*2 (*outer*)
Variables selected	11, 17, 19	8, 11, 14	11, 13, 17	11, 13, 18	5, 11, 13	11, 17, 18
$r_1(1)$	0·733 (1)†	0·915 (1)	0·894 (1)	0·871 (1)	0·917 (1)	0·606 (1)
$r_1(2)$	0·842 (2)	0·236 (1)	0·870 (2)	0·308 (1)	0·726 (2)	0·751 (2)
$r_1(3)$	0·780 (3)	0·949 (2)	0·827 (3)	0·913 (2)	0·719 (3)	0·417 (1)
Q_1	0·751	0·819	0·888	0·792	0·880	0·619

† In brackets beside $r_1(j)$ is the number of the reduced component most highly correlated with the jth original component.

As in the previous example, a direct comparison of components shows the outer-clustering methods to be worse than the others. Here the sets of variables selected by methods B4 and by the inner-clustering version of C2 are best; they have the highest values for Q_1 and a correctly ordered, 1–1 correspondence between original and reduced components.

When the values of the first two reduced components are plotted for each of the 40 aphids, the results are disappointing for all rejection methods. By drawing suitably irregular boundaries, groups which correspond roughly to Jeffers's four groups may be formed but these clusters look very artificial and unconvincing. One major problem is that several of the variables which are retained are discrete and this leads to different aphids having identical values for all three retained variables.

The addition of one extra variable, however, improves the situation considerably. If four variables are retained most of the ties between aphids disappear and the four groups of aphids become fairly well determined. Indeed, for one rejection method, the inner clustering version of C1, the resemblance between the two-dimensional pictures for the full and reduced sets of data is remarkable, as shown in Figs. 1 and 2. Fig. 1 is a rotated version of Jeffers's Fig. 1 with the individual points numbered, and Fig. 2 is a similar two-dimensional plot of the aphids using only the four variables chosen by the inner-clustering version of C1, the single-linkage clustering method.

It must be stressed that none of the other rejection methods give such satisfying pictorial results but if a between-aphids distance matrix is computed using all four retained variables, rather than just two principal components, results are more encouraging. With very few exceptions, aphids belonging to the same Jeffers group

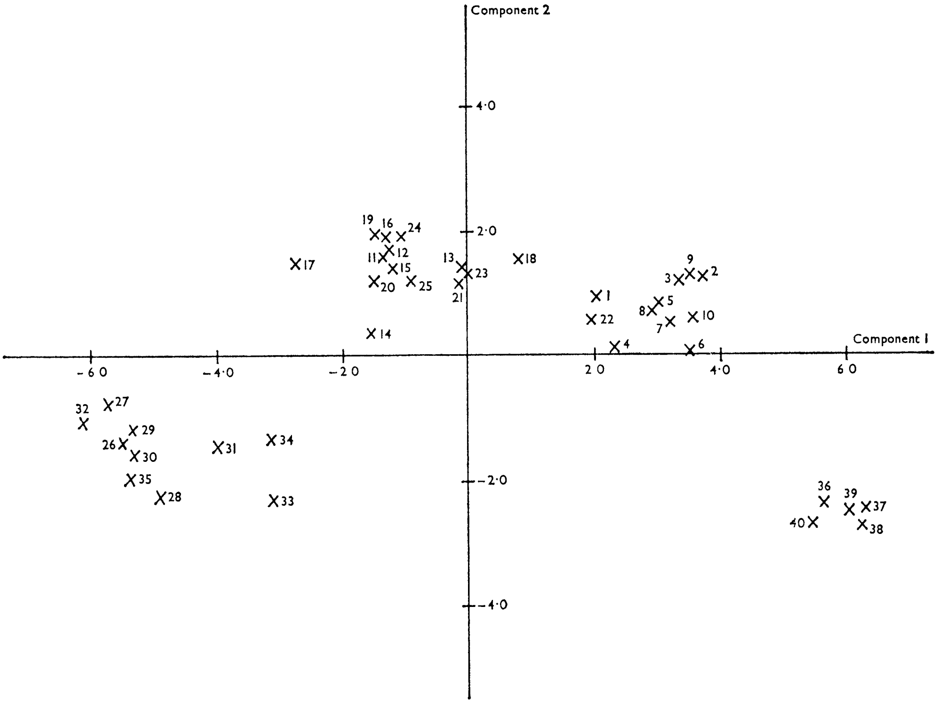

FIG. 1. Plotted values of the first two full components for individual insects.

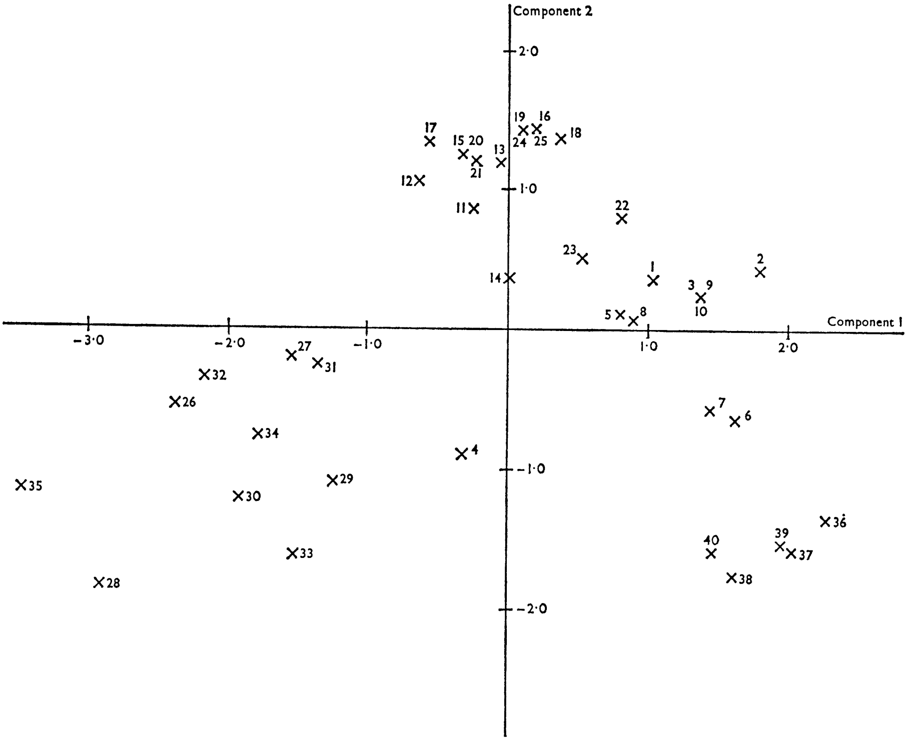

FIG. 2. Plotted values, for individual insects, of the first two components obtained from a reduced set of four variables chosen by the single-linkage inner-clustering method.

have smaller distances between them than aphids from different groups. This is true for all rejection methods. The four-group structure is thus evident when only four variables out of the original 19 variables are retained, and there are very few misclassifications of aphids for any of the rejection methods.

This example shows that the rejection methods may retain too few variables, and that perhaps the values of the rejection criteria chosen empirically from artificial data should be changed slightly. However, when only four of the original 19 variables are retained all the rejection methods give satisfactory results.

Rather than change the rejection criteria, which worked well for the pitprops data, an extra condition, specifying a minimum allowable number of variables to be retained, can be introduced. Certainly two variables are too few for a component analysis, since the components are $(1/\sqrt{2})\ (1,1)$ and $(1/\sqrt{2})\,(1,-1)$, whatever the value of the correlation between the variables. The present data, and the next set, suggest that three variables may also be too few, and that it is advisable to always retain at least four variables.

6. Analysis of Crimes

The data consist of 18 variables, the number of crimes committed in the U.K. within 18 different categories, measured for each of the 14 years 1950–63. Ahamad (1967) does a principal component analysis on these data, and his main conclusion is that the first component, which may be interpreted as a general index of most types of crime, is highly correlated with the population of the U.K. aged 13–18.

Ahamad's method of analysis and the inferences he makes have been criticized (Walker, 1967), but in the present paper the validity of his results is not examined; the question here is whether similar results may be obtained using fewer variables.

It should be mentioned, however, that Ahamad's correlation matrix and the correlation matrix computed from his raw data differ in several entries for variable 16 and in the correlation between variables 3 and 10. In the present paper Ahamad's raw data, rather than his correlation matrix which is not positive semi-definite, have been used.

Of the rejection methods the multiple correlation method suggested three variables should be retained, the average-linkage method suggested four variables were necessary, and the criteria for the other methods were on the border-line between three and four variables. Comparisons were made between components from full and reduced sets of variables retaining three and four variables. The results for four variables were considerably better than for three, so it was decided to retain four variables for all rejection methods.

Subsets of four selected variables, together with values of $r_1(j)$, $j = 1, 2, 3$, and Q^1 based on the first four components, are given in Table 5 for the various rejection methods.

It can be seen from Table 5 that method B2 and the inner-clustering version of C2 do best on a direct comparison of components. Both have high values for all the $r_1(j)$'s, and the full and reduced components are in the same order. The other methods either have lower values for the $r_1(j)$'s, or more than one full component most highly correlated with the same reduced component, or a mixture of both faults.

Ahamad (1967) found that the correlation between the first full component and the population of the U.K. aged 13–18 was 0·989. By comparison the highest correlation between a reduced component and the U.K. population aged 13–18 is given in Table 6 for the various rejection methods.

The values of r in Table 6 are closely related to those of $r_1(1)$ in Table 5 and for each rejection method the reduced component most highly correlated with the U.K. population aged 13–18 and with the first full component is the same. The values of r are fairly high, although most are considerably lower than that for the full set of data. This is not unexpected, however, since the population aged 13–18 was chosen from among other population variables by virtue of its high correlation with the first full component. It is therefore likely that its correlations with other linear combinations of variables, such as the reduced components, would be smaller.

TABLE 5

Measures of similarity, $r_1(j)$, Q_1, between components for the full set of criminal data and for reduced sets of four variables chosen by various rejection methods

Rejection method	*A*2	*B*2	*B*4	*C*1 (*inner*)	*C*1 (*outer*)	*C*2 (*inner*)	*C*2 (*outer*)
Variables selected:	1, 3, 4, 17	1, 7, 10, 13	1, 3, 5, 17	1, 3, 5, 13	1, 3, 13, 17	1, 5, 10, 13	1, 3, 13, 18
$r_1(1)$	0·762 (2)†	0·899 (1)	0·618 (3)	0·929 (2)	0·721 (3)	0·880 (1)	0·883 (2)
$r_1(2)$	0·919 (1)	0·871 (2)	0·788 (1)	0·909 (1)	0·973 (1)	0·863 (2)	0·887 (1)
$r_1(3)$	0·706 (3)	0·902 (3)	0·773 (2)	0·929 (3)	0·972 (2)	0·896 (3)	0·910 (3)
$r_1(4)$	0·670 (4)	0·778 (4)	0·683 (4)	0·372 (1)	0·635 (3)	0·789 (4)	0·365 (1)
Q_1	0·780	0·890	0·656	0·904	0·771	0·875	0·865

† In brackets beside $r_1(j)$ is the number of the reduced component most highly correlated with the jth original component.

TABLE 6

Highest correlation, r, between a reduced component and the U.K. population aged 13–18, for various rejection methods

Rejection method	*A*2	*B*2	*B*4	*C*1 (*inner*)	*C*1 (*outer*)	*C*2 (*inner*)	*C*2 (*outer*)
r	0·776	0·861	0·623	0·918	0·718	0·838	0·847

It is interesting to note that several of the correlations between population aged 13–18 and the original variables are higher than those in Table 6. In fact, variable 7 has a correlation of 0·983 with this population, almost as large a value as for the first full component. Thus Ahamad's main result, that there is a very high correlation between population aged 13–18 and a linear combination of crime variables, can be obtained using only one of the original 18 variables.

7. British Towns

The final set of data is that studied in detail by Moser and Scott (1961), and consists of 57 sociological variables measured for each of the 157 British towns with a population greater than 50,000 in 1951.

Moser and Scott (1961) did a cluster analysis, using principal components, on the data, and produced 16 clusters of towns, two of which contained only one town each, namely London and Huyton with Roby. The clusters formed could be easily interpreted, e.g. one consisted of resorts and spas, one of mainly textile manufacturing towns, one of "exclusive" suburbs and so on. Because of the large size of the present set of data only the quickest rejection method was tested on it. This method, the single-linkage clustering method, successfully reduced the number of variables to fewer than half of the original number, so no other rejection methods were tried.

A comparison was made between the components for the full set of data and for three different subsets of variables produced by the single-linkage rejection method, using the measures $r_2(j)$ and Q_2. These comparisons are summarized in Table 7. The three subsets correspond to values 0·70, 0·60 and 0·50 of the criterion used to decide how many variables to retain. These values are comparable to the 0·55 suggested for artificial data by Jolliffe (1972), and the somewhat higher values in the range 0·60–0·70 necessary for the criminal and alate adelges data above.

Table 7

Measures of similarity, $r_2(j)$, Q_2 between components for the full set of British Towns data and reduced sets of variables chosen by rejection method C1 for various levels of its rejection criterion

Rejection criterion	0·70	0·60	0·50
No. of variables retained	26	17	8
$r_2(1)$	0·705 (1)†	0·683 (3)	0·730 (2)
$r_2(2)$	0·883 (2)	0·662 (1)	0·444 (2)
$r_2(3)$	0·550 (3)	0·562 (4)	0·200 (4)
$r_2(4)$	0·651 (4)	0·478 (2)	0·417 (1)
Q_2	0·702	0·633	0·543

† In brackets beside $r_2(j)$ is the number of the reduced component most highly correlated with the jth original component.

Table 7 shows a steady improvement for higher values of the rejection criterion. At the lowest level the reduced components are mixtures of the full components, at the intermediate level the components are distinguishable but in the wrong order, and at the highest level the components appear in the correct order. Also the value of Q_2 increases as the number of variables retained is increased.

It was therefore decided to use the highest level of the rejection criterion and to retain 26 variables, and a cluster analysis was done on the towns using only 26 variables. This analysis was the same as that used by Moser and Scott (1961), an iterative relocation method using Euclidean distance based on values of the first four principal components. Sixteen clusters were formed in order to give a direct comparison with Moser and Scott's results. Two clusters contained only one town, London and Huyton with Roby, as before. Of the remaining 14 clusters some were very similar to Moser and Scott's but others were different. However, those that were different were not necessarily more difficult to interpret. For example, one of the new clusters contained all four of the similar Midland towns, Derby, Leicester, Nottingham and Northampton, whereas they were split among three of the old clusters. The number of variables in this example can therefore be reduced to only 26 of the original 57 variables, and the clustering of towns obtained is almost as good as for the full set of data. Also, since the rejection method tested is one of the least successful on other data, it may be possible to reduce the number of variables still further using other rejection methods.

8. Discussion

It has been shown that four published examples of principal component analysis use many more variables than are necessary. Similar results could be obtained for all four examples using a half, or fewer, of the original variables.

Also, some of the rejection methods suggested by Jolliffe (1972) have been shown to be suitable for deciding which variables to reject. However, they twice retain too few variables, and on both occasions only two or three variables are retained. It seems reasonable to suggest that in any problem using principal components at least four variables should be retained, and that the rejection methods should be modified to incorporate this condition.

Finally, a comparison between rejection methods yields no firm conclusions. Although the outer-clustering methods are consistently among the worst and the average-linkage inner-clustering method is always among the best, no method is outstandingly good or bad for any of the data.

Further work is necessary, but it seems unlikely that any one method will prove to be overwhelmingly superior to all others.

Acknowledgements

This paper is based, in part, on a D.Phil. Thesis at the University of Sussex, 1970. I am grateful to my supervisor, Professor J. F. Scott, for all his help and advice.

References

Ahamad, B. (1967). An analysis of crimes by the method of principal component analysis. *Appl. Statist.*, **16**, 17–35.

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.

Moser, C. A. and Scott, W. (1961). *British Towns*. Edinburgh: Oliver & Boyd.

Walker, M. A. (1967). Some critical comments on "An analysis of crimes by the method of principal component analysis" by B. Ahamad. *Appl. Statist.*, **16**, 36–39.