Discarding Variables in a Principal Component Analysis. I: Artificial Data
Author(s): I. T. Jolliffe
Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 21, No. 2 (1972), pp. 160-173
Published by: Oxford University Press for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2346488
Accessed: 10-07-2025 21:36 UTC

Discarding Variables in a Principal Component Analysis. I: Artificial Data

By I. T. JOLLIFFE

University of Kent at Canterbury

SUMMARY

Often, results obtained from the use of principal component analysis are little changed if some of the variables involved are discarded beforehand. This paper examines some of the possible methods for deciding which variables to reject and these rejection methods are tested on artificial data containing variables known to be "redundant". It is shown that several of the rejection methods, of differing types, each discard precisely those variables known to be redundant, for all but a few sets of data.

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

1. INTRODUCTION

IN multivariate analysis when a large number of variables, say 10 or more, is available the results are often little changed if a subset of the variables is used. The remaining variables are, in a sense, redundant and can be discarded. Beale *et al.* (1967) deal with the problem of deciding how many variables to discard, and which they should be, mainly with reference to multiple regression analysis. Much other work has also been done on the selection of a best subset of variables in multiple regression, but so far little has been attempted in principal component analysis. Nevertheless, in practical situations the number of variables is sometimes reduced before doing a component analysis. It is certainly useful to reduce the number of variables, if possible, for often variables are present which complicate the data but do not give any extra information. For example, if two of the variables, x_1 and x_2, are such that $x_1 = x_2 + \epsilon$ where ϵ is a random disturbance, then either x_1 or x_2 may be discarded with little loss of information and with little change in the first few principal components. Also, time and money are saved if some of the variables are discarded; computing time is reduced and in future analyses fewer variables need be measured.

Many methods are possible for deciding which variables to reject, but, in practice, experience and intuition often play a part in the selection. Several rejection methods are examined in this paper and compared by testing them on sets of artificial data. The methods have also been tested on some sets of real data, from different fields of application, including data previously used in published examples of component analysis, i.e. Ahamad (1967), criminology; Jeffers (1967), forestry and biology; Moser and Scott (1961), demography. Details of how well the rejection methods work for these real data are given in a forthcoming paper.

Much has been written regarding the relative merits of using either the correlation matrix or the dispersion matrix in a component analysis. Nothing further will be added here but throughout the paper correlation matrices will be used.

2. Rejection Methods

The eight rejection methods discussed in this section can conveniently be divided into three groups. The first consists of two methods which involve multiple correlation coefficients, and which will in future be referred to as Methods A1 and A2. Four methods, B1, B2, B3 and B4, use the principal components in the rejection of variables, and the remaining two methods, C1 and C2, use cluster analysis.

2.1. *Multiple Correlation Methods*

In multiple regression with K independent variables there is an obvious way of choosing a subset of p independent variables. This is to choose the subset which maximizes the multiple correlation of the dependent variable with the p independent variables (or, equivalently, minimizes the residual mean square). An extension of this to the analysis of interdependent variables, of which principal component analysis is a special case, is to retain the set of p variables which maximizes the minimum multiple correlation between the p selected variables and any of the $(K-p)$ rejected variables. This method, hereafter referred to as A1, was suggested by Beale *et al.* (1967). An advantage of A1 is that it finds, in some sense, the best possible subset of p variables, but it is very slow. Beale *et al.* (1967) state that, using the CDC 3200 computer, the method took 15 minutes to apply to 17 variables. The time taken to apply A1 rapidly increases as the number of variables, K, increases and for $K = 30$ several hours of computer time are needed. Method A1 cannot, therefore, be used for large K.

A second method using multiple correlations, A2, is quicker to apply than A1, but the subset of p selected variables is not necessarily the best. Method A2, a stepwise method which seems not to have been previously published, first rejects that variable which has maximum multiple correlation with the remaining $(K-1)$ variables. Then, at each stage, when q variables remain, the variable having the largest multiple correlation with the other $(q-1)$ variables is the next to be rejected. The process continues until p variables remain. One way of deciding upon a suitable value for p is to stop rejecting variables when all multiple correlations between those variables remaining first fall below some R_0. An appropriate value for R_0 can be found by considering the distribution theory of the multiple correlation coefficient, i.e. choose R_0 so that the probability of obtaining a sample multiple correlation coefficient greater than R_0, given that the true multiple correlation is zero, equals α, where α is one of the usual significance levels. However, in practice, values obtained by this means lead to more variables than are necessary being retained.

Although faster than A1, A2 is still fairly slow, e.g. when $K = 30$ it takes about 30 minutes on an ICL 1905 computer, and when $K = 50$, about 4 hours. Most of the computing times given for rejection methods refer to the ICL 1905. However, times for the slowest methods, A1 and B1, are quoted from Beale *et al.* (1967) and refer to the CDC 3200.

2.2. *Methods using the Principal Components Themselves*

As well as Method A1, Beale *et al.* (1967) also mention a rejection method using the principal components themselves, hereafter referred to as Method B1. Here a principal component analysis is performed on all the original K variables, and the eigenvalues are inspected. Then, if p_1 eigenvalues are less than some number, λ_0, the corresponding eigenvectors (the components themselves) are considered in turn, starting with the component corresponding to the smallest eigenvalue, then the

component corresponding to the second smallest eigenvalue and so forth. One variable is then associated with each of these p_1 components, namely that variable which has the largest coefficient in the component (eigenvector) under consideration and which has not already been associated with a previously considered component. The p_1 variables associated with the p_1 considered components are then rejected.

Next, another component analysis is done on the remaining $(K-p_1)$ variables. Again, if any of the eigenvalues are smaller than λ_0 a variable is associated with each of the corresponding components, and these p_2 variables are rejected.

A component analysis is then done on the remaining $(K-p_1-p_2)$ variables, and this procedure continues until all eigenvalues in the latest component analysis are greater than λ_0. At this stage the procedure stops, having reduced the number of variables from K to $(K-p_1-p_2-\ldots-p_l)=p$. The value of p will be determined by the choice of λ_0. Like the methods of the previous section B1 is rather slow. It requires several component analyses to be done with either manual intervention or substantial additional programming between each.

The other methods using the principal components are faster than B1 since each requires only one component analysis to be done, on the full set of K variables. Each of these methods has been used in practice for some time, although nothing has previously been published about them.

Method B2 is the same as the one already described except that only one component analysis is done. If p variables are to be retained, a variable is associated with each of the last $(K-p)$ components, in the same way as before, and these $(K-p)$ variables are rejected.

The next method, B3, again uses the last $(K-p)$ components. For each of the K variables the sum of squares of coefficients of the variable in the last $(K-p)$ components is calculated, and the $(K-p)$ variables for which this sum is largest are rejected. An additional method which turns out to be computationally similar to B3 is to retain those p variables which are best predicted themselves from the first p components. When the correlation matrix is used it can be shown that the proportion of the variance of the ith variable accounted for by the first p components is

$$\sum_{j=1}^{p} \lambda_j a_{ji}^2,$$

where λ_j is the jth eigenvalue, and a_{ji} is the coefficient of the ith variable in the jth component. Thus the method chooses the p variables for which $\sum_{j=1}^{p} \lambda_j a_{ji}^2$ is a maximum, that is for which

$$1-\sum_{j=1}^{p} \lambda_j a_{ji}^2 = \sum_{j=p+1}^{K} \lambda_j a_{ji}^2$$

is a minimum, whereas B3 chooses variables for which $\sum_{j=p+1}^{K} a_{ji}^2$ is minimized. Both methods suffer from the disadvantage that they reject the wrong variables for data of a very simple type, as shown for B3 in Appendix A1 of the present paper. For this reason they are not discussed further.

The last method in this group, B4, involves the use of the first p components and is, in a sense, a backward version of B2. These p components are considered successively, starting with the first, and a variable is associated with each component in the same way as for B1 and B2. These p variables are retained and the remaining $(K-p)$ rejected. In method B1 the number of variables retained, p, is determined by the choice of λ_0, and in the other methods too, p can be chosen to be equal to the

number of eigenvalues of the correlation matrix greater than some λ_0 (with a larger value than for B1). Alternatively, p may be set equal to the number of components needed to account for more than some proportion, α_0, of the total variation, e.g. if $\alpha_0 = 0{\cdot}80$, and if four components account for 0·76, and five for 0·82 of the total variation, then five variables will be retained. It will be shown in Section 3.3 that, in practice, λ_0 appears to be better than α_0 as a criterion for deciding how many variables to reject.

2.3. *Clustering Methods*

Methods of cluster analysis may be used to reduce the number of variables. The K variables are placed in p groups or clusters, and one variable is selected from each cluster. The remaining $(K-p)$ variables are then rejected. Very many clustering methods are possible and the two introduced here are not necessarily superior to other possible types; they are merely examples chosen to test whether, as suggested by Kendall and Stuart (1968, p. 337), cluster analysis can be used successfully in discarding variables. Both methods are of the type discussed extensively in Sokal and Sneath (1963) and both are hierarchical. One method, C1, is a single-linkage method. The other, C2, is an average-linkage method and is similar to the method suggested by Kendall and Stuart (1968, p. 337). Both methods follow the same set of steps, namely:

(a) Define a measure of similarity, r_{XY}, between any two groups of variables X and Y (a single variable is just a special case of a group).
(b) Calculate r_{XY} for each of the $\frac{1}{2}K(K-1)$ pairs of single-variable groups.
(c) If A and B are the two groups for which r_{XY} is a maximum, replace A and B by the single group $C = A \cup B$.
(d) For each group X not involved in the previous amalgamation of groups calculater_{XC}, and return to step (c). The process then cycles through steps (c) and (d) until, if p variables are to be retained, p clusters of variables remain.

In the single-linkage method the measure of similarity between groups X and Y is given by

$$r_{XY} = \max_{\substack{i \in X \\ j \in Y}} r_{ij},$$

where r_{ij} is the correlation coefficient between variables i and j. For the average-linkage method,

$$r_{XY} = \left(\sum_{i \in X} \sum_{j \in Y} r_{ij} \right) \Big/ (n_1 n_2),$$

where n_1, n_2 are the numbers of variables in X and Y respectively.

An obvious way of deciding how many variables should be retained when these methods are used is to continue steps (c) and (d) above until all the r_{XY} between those clusters remaining first fall below some r_0. The process then stops and the number of clusters formed at this stage is the required number of variables. The values of r_0 which are suitable for the two methods of clustering considered here will, of course, be different.

When using clustering methods a procedure is required for selecting one variable from each cluster. Suppose, for example, that a cluster consists of the variables x_1, x_2, x_3, x_4 and that x_1 and x_2 were the first two of these variables to come together.

TABLE 1 *Summary of methods used to discard variables*

Type	*Method*	*Method of selection of p (out of K) variables*	*Criterion for deciding on a value of p*	*Speed*	*Results*
Multiple correlation methods	A1	Selects set of p variables which maximizes the minimum multiple correlation between selected variables and any rejected variable		Very slow (several hours for 30 variables; CDC 3200)	Not tested—too slow
	A2	Step-wise rejection of variables with largest multiple correlation with remaining variables, until p variables remain	Stop rejecting when all multiple correlations between remaining variables first fall below some R_0	Slow (about 30 minutes for 30 variables; ICL 1905)	Always retains good or best set of variables for artificial data. Satisfactory for real data. Suitable value of R_0 is 0·15
Principal component methods	B1	Similar to B2 but reject fewer variables initially, then do another component analysis, reject a few more variables, and so on		Slow (several runs of similar time to B2 with manual intervention or additional computing in between)	Not tested—too slow
	B2	Associate one variable with each of the last $(K-p)$ components, and reject these variables	p equals either (i) the number of eigenvalues of the correlation matrix greater than some λ_0 or (ii) the number of components necessary to account for more than some proportion, α_0, of the total variation	Faster than A1, A2, B1 but slower than C1, C2 (about 30 minutes for 60 variables; ICL 1905)	Retains good or best set of variables for 91 per cent of artificial data. Satisfactory for real data; 0·70 is suitable value for λ_0. No generally acceptable value for α_0
	B3	Associate $(K-p)$ variables with the last $(K-p)$ components and reject these variables			Shown to be unsatisfactory even for data conforming to very simple models
	B4	Associate one variable with each of the first p components and retain these variables			Retains both best and bad subsets more frequently than A2 or B2. Satisfactory for real data; λ_0, α_0: see B2
Clustering methods	C1	Form p clusters of variables using single-linkage cluster analysis and select one variable from each cluster	Keep amalgamating clusters until all measures of similarity between clusters first fall below some r_0	Fast—can be comfortably applied by hand for moderate numbers of variables C1: up to 30–40 variables C2: up to 10–15 variables	Subset of variables depends on how one variable is selected from each cluster, but good or best subset is always retained for the artificial data. C2 is better than C1 for real data, though both are satisfactory. Suitable values of r_0: 0·55 for C1 and 0·45 for C2
	C2	As C1, except use average-linkage cluster analysis			

These were then joined by x_3 and finally by x_4. Among possible ways of selecting one of these variables are:

(i) Choose the last variable to join the cluster, here x_4. This will be referred to as outer-clustering.
(ii) Choose one of the original or inner-most members of the cluster, here x_1 or x_2: inner-clustering.
(iii) Choose one of the four variables at random.

Of the three possibilities, inner clustering is rather better than the other two for the artificial data considered below, and experience with real data suggests that this may be more generally true.

The two clustering methods are much faster than any of the other methods considered. If the correlation matrix is given, both methods can be applied by hand, the single-linkage method for up to 30–40 variables and the average-linkage method for 10–15 variables.

Table 1 summarizes much of the information about the eight rejection methods described above.

3. The Artificial Data

In all, 587 sets of artificial data were generated, conforming to one of five predetermined models (using an ICL 1905 computer). 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. Rejection methods A2, B2, B4, C1 and C2 were tested on the data in order to see whether the variables they rejected were the redundant ones. Methods A1, B1 and B3 were not tested, A1 and B1 because they are too slow, and B3 because, as shown in Appendix A1, it is unsatisfactory.

In Section 3.1 the models will be described, and Section 3.2 will discuss, for each model, which subsets of variables are the "best" to retain, and which, although not "best", are "good" subsets. Section 3.3 will give the results obtained when the rejection methods are applied to the artificial data.

3.1. *The Models*

Five models were used, and in each a number of independent standardized normal variates,

$$z_{ij} \quad (i = 1, 2, \ldots, K; j = 1, 2, \ldots, 100)$$

were generated using a method suggested by Butcher (1960). Linear combinations of the z_{ij},

$$x_{lj} \quad (l = 1, 2, \ldots, K; j = 1, 2, \ldots, 100)$$

were formed and x_{lj} was taken to represent the jth observation on a variable x_l. There are thus 100 observations on each of the variables $x_1, x_2, \ldots, x_K$, and the x_l's are each linear combinations of K independent standardized normal variables $z_1, z_2, \ldots, z_K$.

All five models were constructed so that the variables, x_l, fall into groups; within groups the variables are linear combinations of each other (plus random disturbances), whereas variables from different groups are independent.

In Models I–IV the total number of variables is fixed, as is the number of groups. For the first 3 models there are 6 variables falling into 3 groups and for Model IV 10 variables in 4 groups. In Model V, however, both the total number of variables

and the number of groups may vary, subject to certain constraints, i.e. neither the number of groups nor the number of variables in a group can exceed 10, and the total number of variables cannot exceed 15.

The relationships between the constructed variables, x_l, and the independent standardized normal variates z_i are shown in Table 2 for Models I–IV, but because of its greater complexity Model V cannot be included in the table. Details of the method of construction of Model V are given in Appendix A2.

TABLE 2

Definition of the constructed variables for Models I–IV

Variable	*Model*			
	I	*II*	*III*	*IV*
x_1	z_1	z_1	z_1	z_1
x_2	z_2	z_2	z_2	z_2
x_3	z_3	z_3	z_3	z_2+z_3
x_4	$z_1+0{\cdot}5z_4$	$z_1+0{\cdot}5z_4$	$z_1+0{\cdot}8z_2+0{\cdot}6z_4$	z_4
x_5	$z_2+0{\cdot}7z_5$	$z_2+0{\cdot}7z_6$	$z_2+0{\cdot}7z_5$	$z_4+0{\cdot}75z_5$
x_6	z_3+z_6	z_2+z_6	$z_3+0{\cdot}5z_6$	$2z_4+0{\cdot}75z_5+1{\cdot}5z_6$
x_7	—	—	—	z_7
x_8	—	—	—	$z_7+0{\cdot}5z_8$
x_9	—	—	—	$2z_7+0{\cdot}5z_8+z_9$
x_{10}	—	—	—	$3z_7+z_8+z_9+z_{10}$

3.2. *The Redundant Variables*

For each of Models I–V, $(K-p)$ of the K variables are linear combinations of the other p variables (plus random disturbances). These $(K-p)$ variables are therefore, in a sense, redundant and it was hoped that the rejection methods would discard precisely these variables. There is, however, some choice regarding the best set of rejected variables. For example, in Models I and II, x_4 equals x_1 plus a random disturbance (alternatively x_1 equals x_4 plus a random disturbance) and x_1, x_4 are independent of the other variables. Here x_1 or x_4 should be rejected, but it does not matter which.

Also there may be difficulty in deciding which of several subsets is the best to retain. For example, in Model II, x_5 and x_6 are both x_2 plus (different) random disturbances. Therefore two variables from the group $\{x_2, x_5, x_6\}$ are redundant. Intuitively, it seems preferable to retain x_2 since x_5 and x_6 are more closely related to x_2 than they are to each other. In fact, it can be shown that the optimal variable to retain, in the sense of Method A1, is x_2.

Since the z_{ij} are known to be independently $N(0, 1)$, the population correlation coefficient, ρ_{ij}, between x_i and x_j can easily be found. For Model II

$$\rho_{14} = 0{\cdot}894, \quad \rho_{25} = 0{\cdot}819, \quad \rho_{26} = 0{\cdot}707, \quad \rho_{56} = 0{\cdot}579$$

and all other $\rho_{ij} = 0$. With so many zero values of the correlation coefficients it is fairly simple to calculate, for any set of three selected variables, the minimum value R_m of the multiple correlation coefficient between the selected variables and any of the rejected variables. Method A1 retains the set of variables which maximizes R_m, and

TABLE 3

Subsets of p retained variables classified as "best", "good", "moderate" or "bad", for Models I–IV

Type of subset	*Model*			
	I ($p = 3$)	*II* ($p = 3$)	*III* ($p = 3$)	*IV* ($p = 4$)
Best ($R_m = R_m^*$)	Any subset containing one variable from each of the following groups $\{x_1, x_4\}, \{x_2, x_5\}$ $\{x_3, x_6\}$	$\{x_1, x_2, x_3\}$ $\{x_2, x_3, x_4\}$	$\{x_1, x_2, x_3\}$ $\{x_1, x_2, x_6\}$	Any subset containing one variable from each of the following groups $\{x_1\}, \{x_2, x_3\}, \{x_4, x_5, x_6\}$ $\{x_7, x_8, x_9, x_{10}\}$
Good ($0{\cdot}7R_m^* \leqslant R_m < R_m^*$)	—	$\{x_1, x_3, x_5\}, \{x_1, x_3, x_6\}$ $\{x_3, x_4, x_5\}, \{x_3, x_4, x_6\}$	$\{x_1, x_5, x_6\}, \{x_1, x_3, x_5\}$ $\{x_2, x_4, x_6\}, \{x_2, x_3, x_4\}$ $\{x_3, x_4, x_5\}, \{x_4, x_5, x_6\}$	—
Moderate ($0{\cdot}5R_m^* \leqslant R_m < 0{\cdot}7R_m^*$)	—	—	$\{x_1, x_3, x_4\}$ $\{x_1, x_4, x_6\}$	—
Bad ($R_m < 0{\cdot}5R_m^*$)	All other subsets			

for Model II the appropriate set of variables is either $\{x_1, x_2, x_3\}$ or $\{x_2, x_3, x_4\}$, that is, x_2 is retained rather than x_5 or x_6. If x_2 is replaced in these two sets by x_5 or x_6, then the value of R_m falls from 0·707 to 0·579, and for any other subset of three variables, $R_m = 0$. Thus for Model II, there are two "best" subsets of three variables, and four "good" ones. In general, if R_m^* is the maximum value of R_m, those subsets of p variables for which $R_m = R_m^*$ will be termed "best". "Good" subsets may be defined as those for which $0{\cdot}7R_m^* \leqslant R_m < R_m^*$, and "moderate" subsets as those for which $0{\cdot}5R_m^* \leqslant R_m < 0{\cdot}7R_m^*$. Any subset which has $R_m < 0{\cdot}5R_m^*$ will be termed "bad". With these definitions of best, good, moderate and bad, Table 3 shows which subsets of p variables fall into each category for Models I–IV.

Again, because it is more complicated than the other models, Model V cannot be included in Table 3. In Model V there are K variables which fall into p well-defined groups, the ith group containing m_i variables. Of the variables in the ith group, $(m_i - 1)$ are redundant, so out of the K variables, $(K-p)$ are redundant.

One of the best subsets of p selected variables is $\{x_1, x_2, \ldots, x_p\}$, and the replacement of one or more of these variables, x_i, by another variable from the same group will also give a best subset, for many of the data.

At worst such a subset will still be a good one, but any subset of p variables containing more than one variable from the same group will be bad.

3.3. *Results*

Of the 587 sets of artificial data generated, 96 conformed to each of Models I and II, 95 to Model III, 100 to Model IV and 200 to Model V.

The rejection methods tested were A2, B2, B4, C1 and C2, and, apart from A2, they were tested on all 587 sets of data. A2 was not tested on the data of Model IV, nor was it tested on the original 136 sets of the Model V data. The reason for this was that the program available at the time was too slow. Later, a more efficient program was written for A2 which reduced the time taken considerably, but the raw data of Models IV and V had not been stored, and re-generating them was not thought worth while. Instead an extra 64 sets of Model V data were generated and A2 was tested on these, together with B2, B4, C1 and C2. This was thought to be a reasonable procedure since Model V is more flexible than Model IV, and Model IV is close to being a special case of Model V.

The results of applying rejection methods A2, B2 and B4 to the artificial data are summarized in Table 4, which gives the number and percentage of times each method retains various types of subset of p variables (i.e. best, good, moderate, bad) for Models I–V. The results for C1 and C2 cannot be easily tabulated, and will be dealt with later.

Returning to Table 4, the bottom section gives the overall performance of each of methods A2, B2 and B4. A2 is the most consistent method; it always retains good or best subsets, never moderate or bad ones. B4 retains best subsets more often than A2, but, against this, it retains moderate or bad sets fairly frequently. B2 lies between A2 and B4 in performance. Bad or moderate subsets are retained, though less frequently than for B4, and the proportion of best sets retained lies between those for A2 and B4.

Thus, if the retention of best subsets is considered important, B4 is the appropriate method of the three to use. However, if good subsets are considered only a little worse than the best, both A2 and B2 are superior to B4, and, if time is unimportant,

A2 should be used. If time is important, though, B2 will be preferred, since A2 takes longer to apply than B2 or B4.

The performance of C1 and C2 depends on how a variable is selected from each cluster, but, regardless of how this is done, only good or best subsets are retained for the data of Models I–V, with good sets predominating. This is a similar situation to that for Method A2, although C1 and C2 are much quicker to apply than A2, especially C1. Thus, on the basis of the artificial data, the clustering methods C1 and C2 will be

TABLE 4

Number (and percentage) of times rejection methods A2, B2 and B4 retain various types of subsets of variables for the data of Models I–V

Model	*Type of subset*	*Rejection method*		
		A2	*B2*	*B4*
I	Best	96 (100·0)	96 (100·0)	70 (72·9)
	Bad	0 (0·0)	0 (0·0)	26 (27·1)
II	Best	0 (0·0)	0 (0·0)	91 (94·8)
	Good	96 (100·0)	96 (100·0)	5 (5·2)
	Bad	0 (0·0)	0 (0·0)	0 (0·0)
III	Best	39 (41·1)	73 (76·8)	26 (27·4)
	Good	56 (58·9)	0 (0·0)	5 (5·3)
	Moderate	0 (0·0)	19 (20·0)	63 (66·3)
	Bad	0 (0·0)	3 (3·2)	1 (1·0)
IV	Best	—	96 (96·0)	100 (100·0)
	Bad	—	4 (4·0)	0 (0·0)
V	Best	4 (6·3)	22 (11·0)	180 (90·0)
	Good	60 (93·7)	151 (75·5)	7 (3·5)
	Bad	0 (0·0)	27 (13·5)	13 (6·5)
I–V	Best	139 (39·6)	287 (48·9)	467 (79·6)
	Good	212 (60·4)	247 (42·1)	17 (2·9)
	Moderate	0 (0·0)	19 (3·2)	63 (10·7)
	Bad	0 (0·0)	34 (5·8)	40 (6·8)

preferred to Method A2. There are many possible clustering methods for discarding variables, and there is no reason to suppose that C1 and C2 are the best of these. Despite this they have consistently retained a good or best subset of variables for all the artificial data. It appears, therefore, that clustering methods can be used very satisfactorily in discarding variables.

Regarding the selection of one variable per cluster, inner-clustering (retention of one of the original or inner-most members of a cluster) retains a best subset more often than the other two methods suggested in Section 2.3, for Model II, and for some of the data of Model V. However, for the remaining artificial data it makes no difference how a variable is selected from each cluster.

Finally in this section on results, the artificial data can be used to find suitable values for the various criteria suggested in Section 2 for deciding how many variables to reject. For example, when Method A2 is in use, variables can be rejected one at a time until all multiple correlations between those variables remaining have fallen

below some R_0. If i variables remain, let R_i be the maximum multiple correlation between one of these variables and the other $(i-1)$ variables. Then, if $(K-p)$ variables are known to be redundant, R_0 should be such that

$$R_p < R_0 \leqslant R_{p+1}.$$

For the artificial data p is always known and it was found that if

$$0{\cdot}13 \leqslant R_0 \leqslant 0{\cdot}18$$

then the correct number of variables is always retained, i.e. a suitable value for R_0 is about 0·15 and for the artificial data of Models I–V this value of R_0 always leads to the rejection of the correct number of variables.

In Section 2.2 the criteria λ_0 or α_0 were suggested for deciding how many variables to retain when the principal component methods are used. In order to retain the correct number of variables, p, λ_0 and α_0 should be such that

$$\lambda_{p+1} < \lambda_0 \leqslant \lambda_p, \quad \alpha_{p-1} \leqslant \alpha_0 < \alpha_p,$$

where λ_i is the ith eigenvalue of the correlation matrix, and α_i is the proportion of the total variation accounted for by the first i principal components $(1 \leqslant i \leqslant K)$. There is no value of α_0 suitable for all the artificial data. $\alpha_0 = 0{\cdot}80$ is the best value, but even with this value, the wrong number of variables was rejected on nearly 150 occasions. α_0 is therefore almost useless for deciding how many variables to retain.

Again, with λ_0 no suitable value could be found for all the artificial data. However, if

$$0{\cdot}66 \leqslant \lambda_0 \leqslant 0{\cdot}74,$$

the wrong value of p was given for less than 2 per cent of the data, and the best value for λ_0 was 0·71 for which the wrong number of variables was retained only 5 times out of 587. Furthermore, of these 5 sets of data, 4 were very similar; each consisted of at least 9 groups of variables of which at least 7 contained a single variable. There were many eigenvalues in the range 0·6 to 1·4, and if this occurs it seems advisable to use a somewhat lower value of λ_0, 0·60 say. It seems that usually, though, $\lambda_0 = 0{\cdot}70$ can be used satisfactorily for deciding how many variables to retain.

In the two clustering methods, clusters are combined until all coefficients of similarity between those clusters remaining have fallen below some r_0, and r_0 will differ for the two methods. If $r_0 = 0{\cdot}45$ for the average-linkage method then the correct number of variables is retained for all the artificial data. However, with the single-linkage method there is no universally acceptable value of r_0. If

$$0{\cdot}35 \leqslant r_0 \leqslant 0{\cdot}55,$$

then the correct number of variables is retained for all the data conforming to Models I, II, IV and V, but for Model III a larger value of r_0 was necessary. The best value of r_0 when all the artificial data were considered was between 0·60 and 0·64, and in this range the wrong number of variables was rejected 38 to 40 times (out of 587).

4. Discussion

The results for the various rejection methods have been given in the previous section but it is worth while to give a summary here as well as some results for real data.

Four rejection methods were tested on the artificial data, a multiple correlation method, two principal component methods and two clustering methods. The same

methods have also been tested on several sets of real data, including some used in published examples of principal component analyses, i.e. Ahamad (1967), Jeffers (1967) and Moser and Scott (1961). Details of results for these data will be given in a further paper.

For each rejection method a suitable criterion was found empirically for deciding how many variables to retain. The multiple correlation method works well on the artificial data if variables are discarded until all multiple correlations between remaining variables are less than about 0·15. Similarly, for the principal component methods satisfactory results are obtained if the number of variables rejected equals the number of eigenvalues of the correlation matrix less than about 0·70. Finally, for the clustering methods the appropriate number of variables to retain is the number of clusters when inter-cluster similarities first fall below about 0·55 for the single-linkage method and 0·45 for average-linkage.

The values for these criteria are about right for the real data too, although the values for the clustering methods and, to a lesser extent, the principal component methods may lead to too few variables being retained.

Of the methods tested using these criteria none was notably better or worse than the others for the artificial data. The multiple correlation method and the clustering methods, the slowest and fastest respectively, gave similar results. Both always retained good subsets of variables but only occasionally the best. The principal component methods were more successful at producing best subsets, but bad subsets were also sometimes produced.

For all the sets of real data the number of variables could be reduced by more than a half without appreciably altering the results, and each of the five rejection methods again retained reasonable sets of variables. Also, with one exception, there were no consistent differences in the performances of the methods. The exception was that the average-linkage clustering method was consistently better than the single-linkage method, which is not unexpected since the former is the more sophisticated method. The differences were generally small, though, and single-linkage clustering was still an acceptable method for reducing the number of variables.

Regarding the selection of one variable per cluster in these methods, both real and artificial data show inner-clustering to be superior to outer-clustering. There are, however, occasions when neither should be used. For example, when the variables are discrete a reduction in their number may lead to ties among individuals, and variables should be chosen to prevent this if possible. Again, if one variable in a cluster is much easier to measure than the others, then this one will be chosen.

In conclusion the fact that all the methods give similarly satisfactory results suggests that a method may be selected on consideration of speed alone. Omitting the single-linkage method leaves the average-linkage clustering method as the fastest and therefore most desirable of the methods tested. However, further work is necessary before conclusive results regarding the relative merits of tested methods are available, and it is also possible that other, as yet untested, rejection methods may be superior to those discussed here.

Acknowledgements

This paper is based on part of 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. *Appl. Statist.*, **16**, 17–35.

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

BUTCHER, J. C. (1960). Random sampling from the normal distribution. *Computer J.*, **3**, 251–253.

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

KENDALL, M. G. and STUART, A. (1968). *The Advanced Theory of Statistics*, 2nd ed., Vol. 3. London: Griffin.

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

SOKAL, R. R. and SNEATH, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco: Freeman.

Appendix A1

Demonstration of the Failure of Method B3 on Simple Artificial Data

Suppose the correlation matrix for a particular set of data is of the form

$$\begin{bmatrix} \Lambda_1 & & & 0 \\ & \Lambda_2 & & \\ & & \ddots & \\ 0 & & & \Lambda_p \end{bmatrix},$$

where

$$\Lambda_i = \begin{bmatrix} 1 & \rho_i & \dots & \rho_i \\ \rho_i & 1 & \dots & \rho_i \\ \vdots & \vdots & \vdots & \vdots \\ \rho_i & \rho_i & \dots & 1 \end{bmatrix}$$

is a $(K_i \times K_i)$ matrix,

$$\sum_{i=1}^{p} K_i = K,$$

and all the ρ_i are close to 1 $(1 \leqslant i \leqslant p)$. Then there are p groups of variables, and the ith group contains K_i variables of which $(K_i - 1)$ are redundant. Suppose, without loss of generality, that

$$(K_1 - 1)\,\rho_1 \geqslant (K_2 - 1)\,\rho_2 \geqslant \dots \geqslant (K_p - 1)\,\rho_p.$$

With this permutation of the Λ_i's the largest eigenvalue of the correlation matrix is $1 + (K_1 - 1)\,\rho_1$, the second is $1 + (K_2 - 1)\,\rho_2$, and so forth.

Then, for $1 \leqslant i \leqslant p$, the ith principal component has coefficient $K_i^{-\frac{1}{2}}$ for the K_i variables

$$\left(\sum_{j=1}^{i-1} K_j\right) + 1, \quad \left(\sum_{j=1}^{i-1} K_j\right) + 2, \quad \dots, \quad \sum_{j=1}^{i} K_j,$$

and zero for the remaining $(K - K_i)$ variables.

In Method B3 those $(K - p)$ variables for which S, the sum of squares of coefficients in the last $(K - p)$ components, is largest are rejected. But in the present example S is equal to $1 - (1/K_i)$ for each member of the ith group. It follows that if B3 is used to

reject $(K-p)$ variables then all variables in the uth group will be rejected where u is defined by

$$K_u = \max_{1 \leqslant i \leqslant p} (K_i).$$

Thus the set of variables retained by B3 will not contain one from the uth group and so is unsatisfactory.

Methods B1, B2 and B4, however, each retain precisely one variable from each group, as required.

Appendix A2

Construction of Model V

Suppose, in Model V, that p is the number of groups of variables, m_i is the number of variables in the ith group and $K = \sum_{i=1}^{p} m_i$ is the total number of variables. Then the method of choosing p and the m_i's and constructing the variables $x_1, x_2, \ldots, x_K$ from the independent standardized normal variables $z_1, z_2, \ldots, z_K$ is as follows:

(i) Choose a positive integer, n, in such a way that the probability that $n = i$ is p_i, where

$$p_i = \begin{cases} \frac{1}{4}(\frac{3}{4})^{i-1} & (i = 1, 2, \ldots, 9), \\ 1 - \sum_{j=1}^{9} p_j & (i = 10), \\ 0 & (i = 11, 12, \ldots). \end{cases}$$

(ii) Choose, independently of one another, a further n integers, $m_1, m_2, \ldots, m_n$, in such a way that

$$P(m_q = i) = p_i \quad (i = 1, 2, \ldots; q = 1, 2, \ldots, n).$$

(iii) If

$$\sum_{q=1}^{n} m_q \leqslant 15, \quad \text{put } p = n.$$

Otherwise, define $p < n$ to be such that

$$\sum_{q=1}^{p} m_q \leqslant 15 < \sum_{q=1}^{p+1} m_q.$$

(iv) Generate $100 \sum_{q=1}^{p} m_q$ independent standardized normal variates

$$z_{ij} \ (i = 1, 2, \ldots, \sum_{q=1}^{p} m_q; j = 1, 2, \ldots, 100).$$

(v) Transform the z_{ij} to form x_{lj} as follows:

$$x_{lj} = z_{lj}, \ (l = 1, 2, \ldots, p; j = 1, 2, \ldots, 100),$$

$$x_{lj} = z_{lj} + (1{\cdot}0 + 0{\cdot}2s) z_{sj}, \ \sum_{q=1}^{s-1} (m_q - 1) + p + 1 \leqslant l \leqslant \sum_{q=1}^{s} (m_q - 1) + p,$$

$$1 \leqslant s \leqslant p, \quad 1 \leqslant j \leqslant 100.$$