Step-Wise Clustering Procedures

Author(s): Benjamin King

Source: *Journal of the American Statistical Association*, Mar., 1967, Vol. 62, No. 317 (Mar., 1967), pp. 86-101

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2282912

STEP-WISE CLUSTERING PROCEDURES*

Benjamin King**
The University of Chicago

A simple step-wise procedure for the clustering of variables is described. Two alternative criteria for the merger of groups at each pass are discussed: (1) maximization of the pairwise correlation between the centroids of two groups, and (2) minimization of Wilks' statistic to test the hypothesis of independence between two groups. For a set of sample covariance matrices the step-wise solution for each criterion is compared with the optimal two-group separation of variables found by total enumeration of the possible groupings.

1. INTRODUCTION

In a previously published paper [15] this writer discussed a "quick and dirty" technique for separating a large number of variables into a group of subsets or clusters so that the variables within a cluster would be highly intercorrelated, and variables from different clusters not so highly intercorrelated. No particular originality was claimed for the method, nor was the clustering technique necessarily expected to result in an optimal grouping of the variables into k classes in the sense of minimizing a loss function based on misclassification or some similar argument. Rather, the routine was viewed as a method of exploration properly falling under the heading of "data analysis" rather than "inference" (see Tukey [22]), the results of which would be subject to testing and confirmation via other techniques. The primary virtue of this method is its step-wise nature leading to a simple and rapid computer program involving $n-1$ scannings of a product moment correlation matrix for n variables. At each scanning the variables are classified into a number of groups that is one less than at the previous pass, yielding $n-k$ groups at the kth scanning.

Keeping in mind its rough and ready characteristics, the method was applied to a set of 63 time series of monthly first differences in the logarithm of the price of common stocks. The series were first subjected to a component analysis and an estimated market factor was removed from each observation for each stock. Then the correlation matrix for the residuals was scanned by the cluster technique, and it was hoped that on the 57th pass the six resulting groups would correspond very nearly to the six two-digit Securities and Exchange Commission (SEC) industrial classifications represented in the total set of 63 stocks. As reported in [15], the results were very much in accord with the a priori hypothesis that residual stock price changes cluster according to their industry affiliations. In other words, the six industries—tobaccos, petroleums, metals, rails, utilities, and retail stores appeared to move together over time, and at the 57th pass only three securities were not placed in their proper SEC

* Some of this research was carried out at The RAND Corporation, Santa Monica, California, although the bulk of it was performed with Ford Foundation research funds at the Graduate School of Business, University of Chicago. Any views expressed in this paper are those of the author and should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors.

** Presently a Sloan Faculty Fellow at the Massachusetts Institute of Technology.

groups. At this point more complicated methods of factor analysis were applied to the same residual series and results similar to those of the simple cluster technique were obtained, thus strengthening the hypothesis of the presence of industry factors in the comovement of stock prices.

This technique for grouping variables bears further investigation for two reasons: (1) The factor analytic methods described in [15] entailed the diagonalization of large covariance matrices and the subsequent rotation of factor patterns—both relatively expensive operations on the IBM 7094. The fact that a step-wise clustering procedure using very little computer time led to similar results suggests that in cases where financial resources are scarce the rough and ready method of clustering might alone be a sufficient tool of analysis. The question of the agreement of inferences based on cluster analysis with those based on factor analysis, not to mention the general problem of statistical inference in factor analysis, is, however, outside the scope of this paper. (2) The second reason for studying this method, and that which will be discussed in this paper, is the possibility that through a simple step-wise procedure a researcher can quickly find a distribution of the variables into k groups that is as good as (or almost as good as) the distribution that would be optimal if he had been able to enumerate by brute force the totality of groupings into k subsets and then ranked the groupings on the basis of some criterion of homogeneity or alienation. This brute force enumeration, even when the number of variables is moderate, is very costly, perhaps infeasible, for the largest and fastest computers. For example, the number of ways of separating 25 variables into five (non-empty) groups where the variables within each group are distinguishable is between two and three quadrillion.[1]

Rather than attempt through analysis to prove optimality or near-optimality, we shall content ourselves with the examination of the use of the clustering technique on a set of illustrative correlation matrices gathered from other publications. The cost of total enumeration forces us to deal primarily with the problem of separating the n variables into $k=2$ groups. The criterion according to which the universe of groupings will be ranked for comparison with the result of the quick and dirty method will be the same criterion that determines the new cluster to be formed at each step of the quick and dirty method. In other words we shall see whether or not the step-wise procedure prevents the attainment of the grouping that would be optimal were we free to consider all possible distributions of the variables into two clusters. Two criteria will be examined: (1) the correlation between the sums of the variables in the two groups, described in the next section; and (2) a well-known measure of group alienation developed by S. S. Wilks [24].

2. STEP-WISE CLUSTERING METHOD I

Initially, the n variables are considered to constitute n groups, one variable to each group. The n groups are designated by the subscript i attached to the respective variables ($i=1, 2, \cdots, n$). On the first pass the ($n \times n$) correlation

[1] The number of ways to distribute n distinguishable variables into k non-empty subsets is given by Stirling's number of the second kind [1]. (See also [5], p. 58, prob. 7.)

matrix for the total set of variables is scanned for the maximum value of r_{ij}, the estimated correlation between variable i and variable j. Call

$$\max_{\text{all } i\neq j} r_{ij} = r_{i^*j^*}.$$

When i^* and j^* have been determined the two variables are summed in order to create a new variable. For accounting purposes the new variable is designated $b=\min\{i^*, j^*\}$, and the two merging variables now constitute group b. (The group designated by $m=\max\{i^*, j^*\}$ becomes inactive.) Hence, there are now only $n-1$ groups. In the correlation matrix, the mth row and the mth column are made inactive through the assignment of a low negative number (say -99) to each correlation coefficient, and the bth row and column are replaced by the correlations between the non-merging variables and the new variable b, created by summing variables i^* and j^*.[2]

The second pass involves the scanning of what is effectively an $[(n-1)\times(n-1)]$ correlation matrix. At this pass a third variable may join the group of two variables formed on the first pass. On the other hand, the maximum value of r_{ij} in the revised correlation matrix may again involve two individual variables. Summation of merging variables and revision of the correlation matrix is carried out as in the first pass.

The reader can see that starting with the third pass the two merging groups may both consist of more than one variable, or one group may be an individual variable, or else both merging groups may be individual variables. The last join-up (on the $n-1$st pass) is the trivial one in which all of the variables are clustered into one group.

Summarizing, at each pass the two groups with the highest correlation are merged—hence the number of groups is reduced by one. A cluster of more than one variable is represented by the sum or centroid of the variables, and the correlation matrix is revised at each pass in order to take into account the changes in group content. The aspect of this procedure that leads one to think that optimal groupings may be missed (as in step-wise linear regression) is the fact that after a variable has joined a group of other variables it cannot be removed from that cluster (although the cluster may lose its identity by merging with another cluster on a later pass).

The step-wise clustering method is represented schematically by the flow chart in Figure 1.[3]

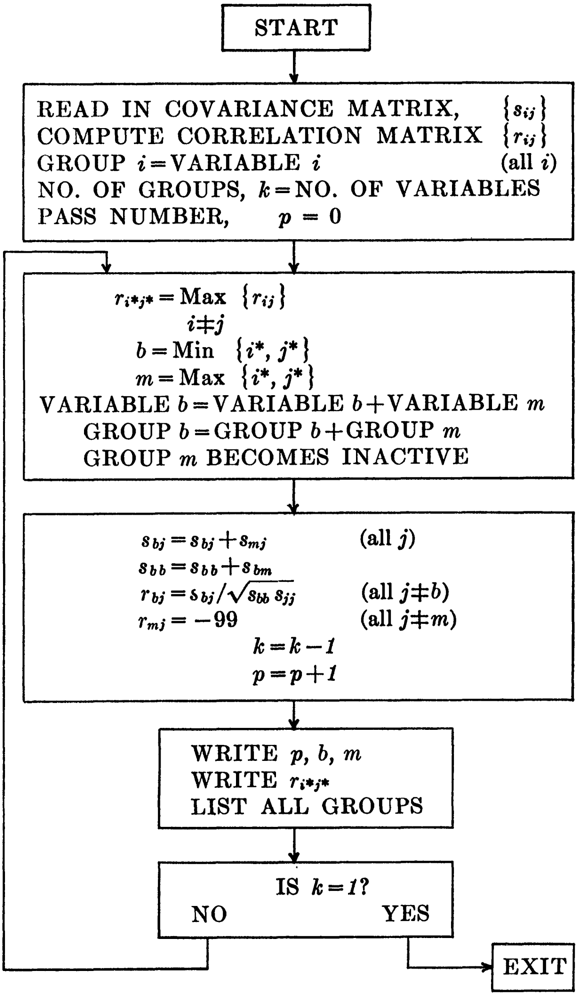
For illustration we have applied the step-wise clustering method to the correlation matrix for six of the common stocks studied in [15]. As explained in Section 1, the basic random variables are first differences in the logarithm of monthly closing price. The series for each stock consists of 403 non-overlapping residual price changes from June, 1927 through December, 1960. The

[2] The revised correlations in row and column b are computed as follows: $r_{bq}=s_{bq}/\sqrt{s_{bb}s_{qq}}=\text{cov}\,(x_{i^*}+x_{j^*}, x_q)/\sqrt{\text{var}\,(x_{i^*}+x_{j^*})\ \text{var}\,(x_q)}=(s_{i^*q}+s_{i^*q})/\sqrt{(s_{i^*i^*}+s_{j^*j^*}+2s_{i^*j^*})s_{qq}}$ where x_{i^*} is variable i^*, and s_{i^*q} is the usual sample covariance. (This operation corresponds to that shown in the flow chart, Figure 1.)

[3] In [15] we have recognized the fact that this clustering technique is a special case of "hierarchical grouping to optimize an objective function" discussed by Ward [23]. The objective function in this case is merely the pairwise correlation coefficient between group centroids.

FIGURE 1

FLOW CHART OF CLUSTER METHOD I


START
READ IN COVARIANCE MATRIX, $\{s_{ij}\}$
COMPUTE CORRELATION MATRIX $\{r_{ij}\}$
GROUP i = VARIABLE i (all i)
NO. OF GROUPS, k = NO. OF VARIABLES
PASS NUMBER, $p = 0$
$r_{i^*j^*} = \underset{i \neq j}{\text{Max}} \{r_{ij}\}$
$b = \text{Min} \{i^*, j^*\}$
$m = \text{Max} \{i^*, j^*\}$
VARIABLE b = VARIABLE b + VARIABLE m
GROUP b = GROUP b + GROUP m
GROUP m BECOMES INACTIVE
$s_{bj} = s_{bj} + s_{mj}$ (all j)
$s_{bb} = s_{bb} + s_{bm}$
$r_{bj} = s_{bj}/\sqrt{s_{bb}\, s_{jj}}$ (all $j \neq b$)
$r_{mj} = -99$ (all $j \neq m$)
$k = k - 1$
$p = p + 1$
WRITE p, b, m
WRITE $r_{i^*j^*}$
LIST ALL GROUPS
IS $k = 1$?
NO
YES
EXIT

securities and their estimated residual correlations are displayed in Table 1. There are three stocks each from the tobacco and the petroleum industries respectively.

TABLE 1. CORRELATION MATRIX FOR RESIDUAL PRICE CHANGES, THREE TOBACCO STOCKS AND THREE PETROLEUM STOCKS

Security	2	2	3	4	5	6
American Tobacco	1.000					
Liggett and Myers	0.776	1.000				
Reynolds Tobacco	0.799	0·761	1.000			
Continental Oil	−0.220	−0.150	0.077	1.000		
Atlantic Refining Co.	−0.242	−0.188	0.077	0.819	1.000	
Skelly Oil	−0.381	−0.236	0.030	0.780	0·805	1.000

In Figure 2 the operation of the clustering technique can be observed by covering the chart with a piece of paper and then exposing a column at a time from left to right. On the first pass the maximum correlation is 0.8194, between Continental Oil and Atlantic Refining. Thus we have the formation of the first group of two stocks, indicated by the black bars in column 1.

On the second pass, Variable 6, Skelly Oil, joins the previously merged petroleums. The correlation between Skelly Oil and the sum of Continental and Atlantic is shown in the bottom row to be 0.8295. (In order to verify this we need, of course, the covariance matrix.) At this stage we now have four clusters of variables—three consisting of one variable each and a three-variable petroleum group.

On subsequent passes, the tobaccos join together (indicated by the cross-hatched design), and finally at pass 4 we have two distinct clusters of three variables each.

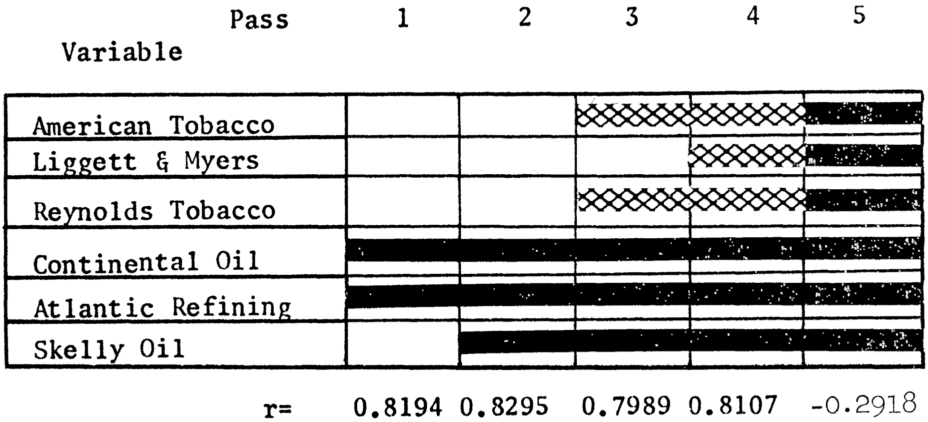


FIG. 2—Step-wise cluster analysis of common stock price changes after removal of first centroid.

On pass 5 the final merger occurs with a correlation of -0.2918 between the two groups' centroids. Next we ask whether there is any other grouping of the six variables into two clusters that has a lower (more negative) between-group correlation, and since the number of variables is small, we can check for the optimal grouping by total enumeration.[4] In Table 2, the composition of a group is indicated by a string of zeros and ones in column positions corresponding to the six variables. Alongside the composition indicator is the value of the correlation coefficient between the sums of the variables in each group. We see that the lowest correlation in the rank is indeed the grouping that was determined by the step-wise procedure.

In principle, one could perform a similar check upon the optimality of the selection of $n - k$ clusters at any pass k, but as we have already mentioned, we shall restrict our primary interest in this paper to the two groups formed on the next to the last pass.

[4] Stirling's number of the second kind for $k=2$ groups and n variables is equal to $2^{n-1}-1$.

TABLE 2. TOTAL ENUMERATION OF GROUPINGS FOR COMMON STOCKS

Cluster Pattern	r Value
000111	−0.291777
010111	−0.123789
001111	−0.110287
011000	−0.037290
011111	0.040846
000101	0.057638
000011	0.109352
001000	0.118351
010000	0.163974
000001	0.334272
000110	0.340834
010101	0.406600
001101	0.414890
001011	0.469841
010011	0.482195
011010	0.529629
000100	0.558944
000010	0.587284
011100	0.590937
001110	0.673922
010001	0.690517
001001	0.715106
011001	0.719681
011101	0.721329
010110	0.729551
011011	0.743687
001010	0.776659
011110	0.789473
001100	0.794033
010100	0.802300
010010	0.816726

In the particular form of the step-wise clustering procedure used in the example above, the criterion for merger at each pass was that of maximum correlation between group centroids. The result is the determination of the two groups that move most counter to one another, and when considerable negative covariance is present, there may be fairly high negative correlation between the final pair of clusters. If one objects to this result, and desires instead a pair of clusters with small intercorrelation (of either sign) he can easily substitute r^2 for r in the flow chart in Figure 1 and run the program accordingly. It is the consideration of the problem of finding clusters that are as uncorrelated as possible that led us to the next form of the step-wise procedure.

3. STEP-WISE CLUSTERING METHOD II

Over thirty years ago S. S. Wilks [24] developed a likelihood ratio test for the hypothesis that k subsets of n multivariate normal variables are mutually independent. The essential statistic is

$$W = |A| \Big/ \prod_{i=1}^{k} |A_{ii}| \tag{3.1}$$

where $|A|$ is the determinant of the $n \times n$ matrix of sample crossproduct deviations for the full set of variables, and $|A_{ii}|$ is the determinant of the matrix of crossproduct deviations for the ith subset. (See Anderson [2] for a full discussion of the problem.)

There is a well-known geometric interpretation of the statistic W that bears repeating because it is so intuitively appealing: let the T observations of the n variables (in terms of deviations from their arithmetic means) be represented by the columns of a $T \times n$ matrix X. Then the matrix of sample sums of crossproducts, A, is given by $X'X$. The determinant $|A|$ is the squared volume of the parallelotope with the columns of X as principal edges. Now, if the k subsets of the n variables are in fact mutually uncorrelated, the subparellelotopes whose squared volumes are represented by $|A_{ii}|$, $i = 1, \cdots, k$, are orthogonal to one another. Hence the squared volume $|A|$ is given by the product

$$\prod_{i=1}^{k} |A_{ii}|,$$

and $W = 1$. The greater the collinearity among the subparallelotopes, the greater is the denominator of W relative to the numerator, and the closer is W to zero. Therefore, quite apart from its value as a test statistic, the ratio of determinants, W, serves nicely as an index number to indicate the degree of togetherness of k subgroups of variables on the basis of their sample intercorrelations. W is easily shown to be invariant under all diagonal linear transformations of the matrix X, hence one may work with sample cross products, covariances, or correlation coefficients according to taste.

In the case of two subgroups with one of the groups consisting of one variable, and the other containing $n-1$ variables, the statistic W reduces to $1-R^2$, where R^2 is the coefficient of multiple determination for the regression of the single variable on the remaining $n-1$. Rozeboom [18] has suggested the use of Wilks' statistic as a measure of multivariate association, calling it the square of the "generalized alienation coefficient," and he shows that when k equals 2,

$$W = |A| / |A_{11}| \cdot |A_{22}| = \prod_{i=1}^{m} (1 - r_i^2), \tag{3.2}$$

where r_i is the ith canonical correlation between the two sets of variables. (When the smallest subgroup consists of one variable, there is, of course, only one canonical correlation, namely R.) This topic and related ideas have been exposited in an econometric context by J. W. Hooper [10]. One must also mention the pioneering work of Hotelling in this area [11].

With the IBM 7094 it is an easy matter to insert a matrix inversion routine into the program in Figure 1, calculate determinants of submatrices, and at each pass search for the minimum two-group value of Wilks' statistic, i.e., the maximum squared generalized correlation, instead of the relatively crude correlation between centroids.

TABLE 3[5]. CORRELATION MATRIX FOR FIVE TRAITS OF SCHOOL CHILDREN

Trait	1	2	3	4	5
Arithmetic Speed	1.0000				
Arithmetic Power	0.4249	1.0000			
Intellectual Interest	−0.0552	−0.0416	1.0000		
Social Interest	−0.0031	0.0495	0.7474	1.0000	
Activity Interest	0.1927	0.0687	0.1691	0.2653	1.0000

We illustrate this modified step-wise procedure with the correlation matrix that Wilks employed in his 1935 article [24, 13] (Table 3).

With Figure 3 one can follow the progression of the step-wise clustering program. The value at the bottom of each column is now the minimum W at each corresponding pass. The final grouping of variables 1, 2 versus 3, 4, 5 is the same pattern for which Wilks tested the hypothesis of zero group correlation: and the final value of W, 0.9422, agrees with his calculation.

Table 4 shows that the result of the step-wise method is also the best grouping that can be obtained.

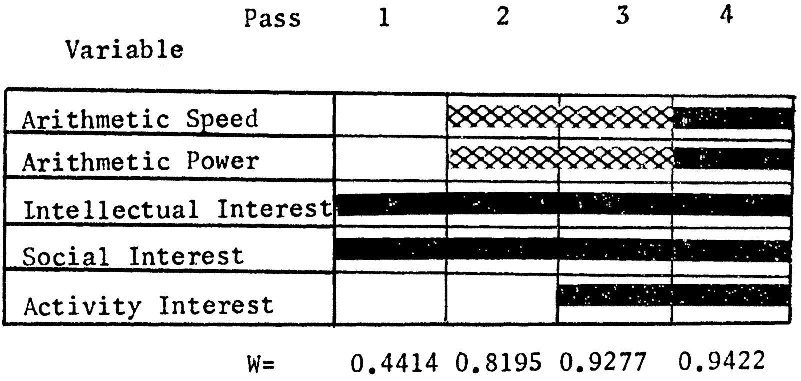


Fig. 3—Step-wise cluster analysis of traits of school children (Method II).

4. SOME EMPIRICAL RESULTS

From sociological and psychological publications we have chosen five correlation matrices on which to try the two techniques of clustering discussed in the previous sections of this paper. In each case, after Methods I and II were applied, the criterion values for the $2^{n-1}-1$ possible two-cluster patterns were computed and ranked as in Tables 2 and 4.

[5] It is interesting to note Wilks' comment [24 p. 324] in referring to the study by Kelley [13] from which these correlations were obtained. He mentions that raw correlations were used because the correlations corrected for attenuation can, among other objections, lead to a negative determinant for the correlation matrix. We have observed the same troublesome phenomenon when tetrachoric r is employed as an estimate of the product moment correlation coefficient. One can construct examples of pairwise dichotomous frequency distributions for three variables for which the matrix of tetrachoric coefficients (read from the Chesire, Saffir and Thurstone Computing diagrams [4]) has a negative determinant. This is most likely due to the skewness and kurtosis of the underlying processes, since the tetrachoric estimates are based on the bivariate normal distribution function. [See 14, 4 and 16].

TABLE 4. TOTAL ENUMERATION OF GROUPINGS FOR TRAITS OF SCHOOL CHILDREN

Cluster Pattern	W Value
01101	0.327983
01010	0.341674
01011	0.343127
01100	0.355002
00101	0.398443
00010	0.416149
00011	0.416475
00100	0.434257
01001	0.724291
01110	0.756361
01111	0.787234
01000	0.808746
00001	0.890299
00110	0.908020
00111	0.942248

A summary of our results is presented in Table 5. For example, when the correlation matrix of order 10 from reference [17] is subjected to Method I, the step-wise result for $k=2$ is variables (1, 5, 7, 9, 10) versus the remainder. The correlation between the group centroids is 0.2622. There are, however, two better groupings, as shown by the total enumeration of the 511 possible groupings. The optimal pattern is (1, 5, 10) versus the others, with a minimum r of 0.2581. We see that, although the step-wise solution is not optimal, it is close to the best possible clustering, if one accepts a difference of .0041 in the correlation coefficients as "close". With Method II, on the other hand, both the step-wise procedure and the total enumeration give the same result, variable 9 versus the others, with a maximum W of 0.6642.

The second method, based on Wilks' statistic, is no more successful than the matrix inversion routine that is used to compute the required determinants. That is, if the correlation or covariance matrix under analysis is very close to singular, truncation error may give anomalous results. Either for this reason, or because of typographical errors in the journals, several matrices that we had planned to examine appeared to have negative determinants. (Since it is a common practice to report only the upper or lower triangle of correlation matrices, one cannot even check for symmetry.) Furthermore, it is possible to get a negative determinant if tetrachoric correlation estimates are used. (See footnote 5, above.) An example of the latter is the (19×19) tetrachoric matrix in a paper by Kahl and Davis [12]. We were especially interested in this matrix because it was used by Fortier and Solomon [6] in order to compare the results of Tryon's method of cluster analysis [21] with a new technique that Fortier and Solomon call the C^* method.[6]

[6] The Fortier and Solomon clustering procedure requires that for each pair of variables i and j, a quantity $D_{ij}=(\rho_{ij}-.5)$ be computed. ρ_{ij} is the zero order sample correlation and .5 is an arbitrary criterion. It is shown that a total gain function can be maximized if pairs are selected by considering only positive values of D_{ij}. The optimal number of pairs, p, and then the optimal selection of p D_{ij}'s out of the total number of positive D_{ij}'s is determined by total inspection, which, although involving fewer quantities than Stirling's number of the second kind, can require much computer time. Fortier and Solomon suggest that in large-scale problems sampling methods may be used in order to find the optimal p and the optimal set of p pairs to be used as the nuclei for the clusters (see [6] for precise details).

TABLE 5. RESULTS OF TEST OF CLUSTER METHODS I AND II ON FIVE SAMPLE CORRELATION MATRICES

Source[a]	No. of Variables	No. of Possible Groupings	Method I		Method II	
			Step-wise (Group 1) Min r No. Better	Total (Group 1) Min r	Step-wise (Group 1) Max W No. Better	Total (Group 1) Max W
[19, p. 65]	9	255	(1, 2, 3, 4) −0.2658 2	(1, 2, 3) −0.3223	(1) 0.8478 0	(1) 0.8478
[9]	10	511	(2) −0.0575 0	(2) −0.0575	(3) 0.8489 2	(2) 0.8632
[17, p. 461]	10	511	(1, 5, 7, 9, 10) 0.2622 2	(1, 5, 10) 0.2581	(9) 0.6642 0	(9) 0.6642
[20, p. 160]	12	2047	(1, 10) 0.6336 1	(1) 0.5505	(1) 0.4444 0	(1) 0.4444
[8, p. 137][b]	15	16383	(10, 11, 12, 13 −0.0358 1	(10, 11, 12, 13, 14, 15) −0.0472	(14, 15) 0.9244 1	(2) 0.9377

[a] Numbers correspond to sources in list of references.
[b] The first 15 variables from a total of 24 are used here.

Although the use of Method II was precluded we were able to apply Method I and compare it to the Tryon and C^* results for the Kahl and Davis data. The first three columns of Table 6 are taken from Fortier and Solomon. The second column shows the eight-cluster pattern obtained by Kahl and Davis using Tryon's technique. Since variable 5, "North-Hatt scale occupation" and variable 7, "Interviewer's rating of subject," are omitted, it is presumed that they did not appear to cluster with any other variables. In the third column are the results of the Fortier and Solomon C^* method. "Not a cluster" means that the variables remained separate from all others (presumably 5 and 7 as well). Finally in the fourth column we have added the results of Method I, halted on the ninth pass in order to produce ten clusters.[7] The complete Method I solution for all values of k is shown in Figure 4.

Table 6 shows that our Method I solution for ten clusters is similar to the Kahl and Davis results except that variables 3 and 8 are merged with 1, 4 and 10. Furthermore, in our results variables 5 and 7, rather than sitting alone, are part of that cluster also. The clustering of subject's education and self-

[7] In our method we consider even a lone variable to constitute a cluster. Hence, by that definition, with the addition of variables 5 and 7, Kahl and Davis would have ten clusters.

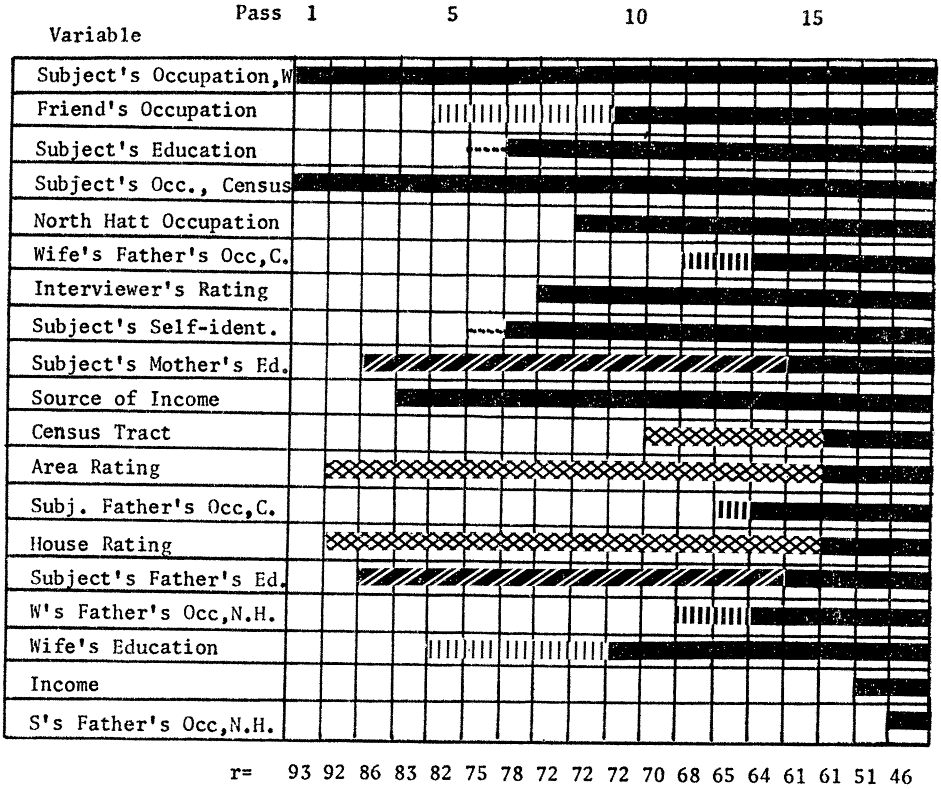


FIG. 4—Method I cluster analysis of Kahl and Davis data.

identification with occupation and income would not seem unreasonable a priori. Note, however, that if we apply our definition of a cluster as "one or more variables" to the C^* solution in Table 6, we find that (counting single variables 5 and 7) Fortier and Solomon have thirteen clusters. This corresponds to Pass 6 for Method I; and examination of Figure 4 shows *exactly the same groupings* as the C^* method.

We must remark that the Tryon and the C^* methods are concerned with determination of the optimal number of clusters as well as the optimal pattern for that number. Our step-wise procedures merely provide solutions for all numbers of clusters without choosing an optimum value of k.

Finally, we mention another somewhat undesirable phenomenon that occurs in Method II: In Table 5 each of the optimal two-cluster patterns under Method II consists of one variable versus the remaining $n-1$ variables. Were it not for the five variable examples in Wilks' original paper [24], and two more examples to be reported below, we might think that the optimal solution for $k=2$ *always* involves one variable against the remainder. We believe, however, that this phenomenon is more likely to be found in covariance matrices whose determinants are not far from zero as is unfortunately the case with the examples that we have chosen. Hooper [10] has commented indirectly on this

TABLE 6. COMPARISON OF RESULTS OF TRYON'S METHOD, C* METHOD, AND METHOD I OF THIS PAPER WHEN APPLIED TO KAHL AND DAVIS DATA

Clusters[a]	Tryon's Method	C* Method	Method I
	Variables		
1	12. Area rating 14. House rating	The same variables	The same variables
2	15. Subject's father's education 9. Subject's mother's education	The same variables	The same variables
3	2. Friend's occupation 17. Wife's education	The same variables	The same variables
4	4. Subject's occupation, Census 1. Subject's occupation, Warner 10. Source of income	The same variables	The same variables merged with variables 3, 8, 5 and 7
5	16. Wife's father's occupation, North-Hatt 6. Wife's father's occupation, Census	Not a cluster	Not a cluster
6	11. Census tract 18. Income	Not a cluster	Not a cluster
7	3. Subject's education 8. Subject's self-identification	The same variables	See Cluster 4 above
8	19. Subject's father's occupation, North-Hatt 13. Subject's father's occupation, Census	Not a cluster	Not a cluster

[a] The clusters are ordered by decreasing values of Tryon's index.

matter in rejecting the generalized alienation coefficient, W, as a measure of the association between two sets of variables in simultaneous econometric equation systems. He observes that because W in the case of $k=2$ is the product of the complements of the squared canonical correlations (see 3.2), when the number of canonical correlations is large, the statistic tends to zero as the result of the introduction of multiplicative factors that are less than one and decreasingly positive. This might partly explain the higher values of W for the case where the number of canonical correlations is only one, but there does not appear to be any obvious analytical reason why a solution with two or more canonical correlations cannot lead to a higher W in some sets of variables. One way out, suggested by Hooper is to use the geometric mean canonical alienation, i.e., the mth root of W, where m is the number of variables in the smaller cluster of the two. He finally chooses a related measure which he calls the "trace correlation".

Any of these alternative measures of group alienation can be easily substituted into the step-wise program, but we shall not explore the topic further at this time.[8]

As an additional test of the performance of the two step-wise methods in the two group case, the following "restricted" Monte Carlo approach was employed. From the (63×63) residual covariance matrix for security price changes, referred to in the introduction and in [15], a succession of random principal submatrices of order 10 was chosen, and the two clustering methods applied to each submatrix. Following each step-wise analysis, the total enumeration for that sample of variables was examined for the optimal grouping, and the discrepancies recorded.[9] We realize that this is not as thorough a test as one based on the unrestricted random generation of covariance matrices; but it was decided that within the (63×63) parent matrix representing six industrial classifications there was sufficient looseness of the factor structure to result in a variety of situations in the sampled (10×10) submatrices.

The results of the experiment for Method I are displayed in Figure 5. The histogram illustrates the frequency distribution for the number of cluster patterns that are better than the step-wise solution. The maximum observed number of better clusters was 12 and this occurred in only one sample submatrix. Since for 10 variables in two groups, Stirling's number of the second kind is 511, our result indicates that 0.023 of the total set of possible cluster patterns was better than the step-wise result for that sample. For each frequency category in Figure 5 we display the maximum observed absolute difference between the step-wise intercluster correlation and the correlation for the optimal grouping. We see that even in the sample with 12 better solutions, this absolute difference is only 0.0943; although, in the one case of 10 better solutions, the difference is 0.2012.

With Method II applied to the same sample submatrices, 38 out of the 45 cases resulted in a step-wise solution that was also optimal. In the remaining 7 instances only one grouping was preferred to the step-wise result and the maximum absolute difference between the two values of W was 0.0745, with a mean absolute difference of 0.0236. As expected, only 2 of the 45 cases led to a solution involving more than one variable in the smaller cluster. In one case the optimal grouping consisted of 2 variables versus the remaining 8 and the step-wise procedure captured it. In the second example, the best grouping con-

[8] The interested reader should see the recent paper by Friedman and Rubin [7] in which several competing cluster criteria based on multivariate test statistics are examined and applied to sets of data. Remark, however, that the discussion in [7] centers on the clustering of a sample of vector observations into homogeneous sub-groups, rather than the clustering of variables.

[9] A programming error destroyed the 63rd diagonal element, so that the actual sampling was from a (62×62) covariance matrix. The total number of possible submatrices is $\binom{62}{10}$ from which we have randomly selected 45.

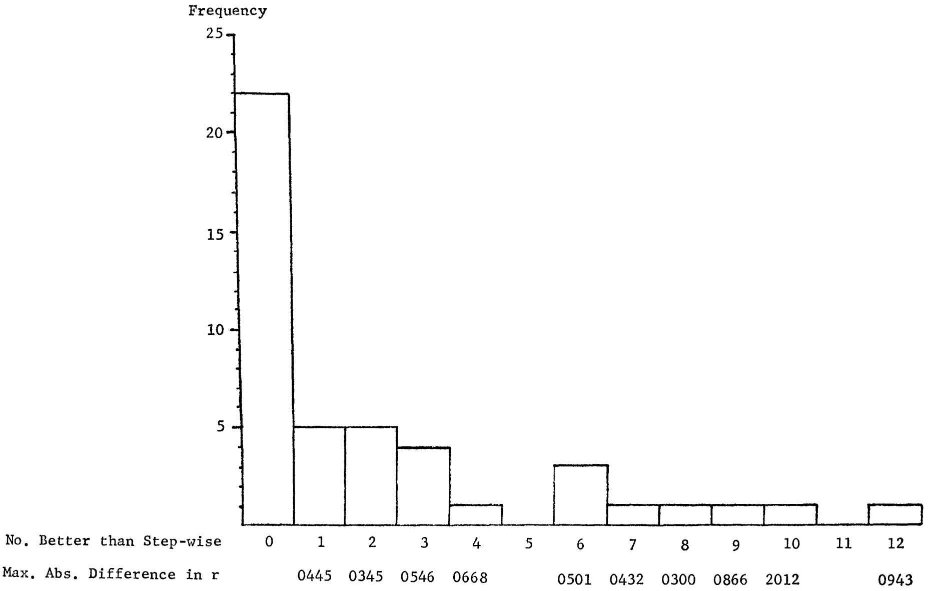


FIG. 5—Frequency distribution of number of groups better than step-wise solution in sampling experiment for Method I.

sisted of 3 variables in the smaller group, but the step-wise procedure chose the second best cluster pattern according to the ranked values of W.

5. CONCLUSION

To conclude our discussion of step-wise techniques and empirical checks, we observe that in less than half of the trials of Method I do we obtain the optimal separation into two groups. Nevertheless, we are seldom very far from the best solution whether we measure the discrepancy in terms of the number of solutions that are better than the step-wise result, or in terms of the difference in the between centroid correlations for the step-wise solution and the best solution.

With Method II, agreement between the step-wise result and the grouping of highest rank appears to be more frequent.

Note, however, that neither the number of better groupings, nor the difference in criterion values between the step-wise and the optimal solutions are completely adequate measures of performance for the evaluation of step-wise techniques. For example, the step-wise solution may be second in rank, and yet the optimal result may have a criterion value that is far greater than that for the step-wise result. On the other hand, even when the step-wise and the optimal groupings are very close with respect to their criterion values, it may be difficult to translate the difference into meaningful concepts of correlation and alienation; e.g., "How different with respect to the separation of variables are two groupings whose criterion values in Method I differ by 0.0041?" In the third row of Table 5 we see that with an observed criterion difference of 0.0041, the step-wise solution differs from the optimal grouping by two variables. (We have a similar situation in the last row of the table.) In reply to these possible points of criticism, we remark that except in situations where subsets of variables are distinctly different from one another we can only expect step-wise methods to carry us within a neighborhood of the optimal result. But at that point the researcher who is familiar with the variables, and who doubtless has prior feelings about the way that they should be grouped, can either by hand or with the computer make minor exchanges of variables and observe whether or not the criterion function is increased. The reader may be familiar with certain programs for step-wise linear regression in which at each step there is some provision for the release of variables from the equation as well as the addition of variables. Doubtless, similar modifications could be made in the program illustrated in Figure 1 at the expense of an increase in running time, but without approaching the time required for complete enumeration.

Finally, we do not want to imply that the fact that the step-wise procedure appears to perform better in one method than in another means that the corresponding 'criterion for clustering is to be preferred. The list of clustering techniques compiled by Ball [3] shows that there are a great number of competing criteria, each of which has its merits and shortcomings, depending on one's definition of homogeneity and the model that he has in mind. At the level of data analysis the choice of a model may be quite subjective, with the expectation that additional observations and predictive tests will either reinforce or cast doubt upon that choice.

6. ACKNOWLEDGMENTS

I should like to thank the referees and the editor for suggestions that have helped to improve the presentation of my results. In addition, I am grateful to Gary Eppen, Lawrence Fisher, Harry Roberts, William Sharpe and Herbert Solomon for non-binding discussion of several problems related to this subject.

REFERENCES

[1] Abramowitz, M. and Stegun, I. A. (eds.), *Handbook of Mathematical Functions*. Washington, D.C.: National Bureau of Standards, (1964).

[2] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*. New York; John Wiley and Sons, (1958).

[3] Ball, G. H., "Data Analysis in the Social Sciences: What about the Details?", *Proceedings—Fall Joint Computer Conference*, (1965)

[4] Chesire, L., Saffir, M., and Thurstone, L. L., *Computing Diagrams for the Tetrachoric Correlation Coefficient*. Chicago: University of Chicago, (1933).

[5] Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. I. New York: John Wiley and Sons, Inc., (1957).

[6] Fortier, J. J., and Solomon, H., "Clustering Procedures," *Proceedings of the International Symposium on Multivariate Analysis*. New York: Academic Press, (1966)

[7] Friedman, H. P., and Rubin, J., "On Some Invariant Criteria for Grouping Data," *IBM New York Scientific Center Technical Report 39.001*, (1966).

[8] Harman, H., *Modern Factor Analysis*. Chicago: University of Chicago Press, (1960).

[9] Holzinger, K., and Swineford, F., "A Study in Factor Analysis: The Stability of a Bi-factor Solution," *Supplementary Educational Monographs, No. 48*. Chicago: University of Chicago, (1939)

[10] Hooper, J. W., "Simultaneous Equations and Canonical Correlation Theory," *Econometrica*, 27 (1959), 245–56.

[11] Hotelling, H., "Relations between Two Sets of Variates," *Biometrika*, 28 (1936), 231–77.

[12] Kahl, J. A., and Davis, J. A., "A Comparison of Indexes of Socioeconomic Status," *American Sociological Review*, 20 (1955), 317–25.

[13] Kelley, T. L., *Crossroads in the Mind of Man*. Stanford: Stanford University Press, (1928).

[14] Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics*. London: Charles Griffin and Co., Ltd., (1963).

[15] King, B. F., "Market and Industry Factors in Stock Price Behavior," *Journal of Business*, 39, Part II (1966), 139–90.

[16] Newbold, E., "Notes on an Experimental Test of Errors in Partial Correlation Coefficients Derived from Fourfold and Biserial Total Coefficients," *Biometrika*, 17 (1925), 251–65.

[17] Rinn, J. L., "Structure of Phenomenal Domains," *Psychological Review*, 72 (1965).

[18] Rozeboom, W. W., "Linear Correlations between Sets of Variables," *Psychometrika*, 30 (1965), 57–71.

[19] Short, J. F., Jr., Rivera, R., and Tennyson, R. A., "Perceived Opportunities, Gang Membership, and Delinquency," *American Sociological Review*, 30 (1965), 56–67.

[20] Thurstone, L. L., *Multiple Factor Analysis*. Chicago: University of Chicago Press, (1947).

[21] Tryon, R. C., *Cluster Analysis*. Ann Arbor: Edwards Bros., (1939).

[22] Tukey, J. W., "The Future of Data Analysis," *Annals of Mathematical Statistics*, 33 (1962), 1–67.

[23] Ward, J. H., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58 (1963), 236–44.

[24] Wilks, S. S., "On the Independence of k Sets of Normally Distributed Statistical Variables," *Econometrica*, 3 (1935), 309–26.