DOCUMENT RESUME

ED 065 535 TM 001 457

AUTHOR	Harman, Harry H.
TITLE	How Factor Analysis Can Be Used in Classification.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPCRT NO	ETS-RB-71-65
PUB DATE	Dec 71
NOTE	18p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Classification; Cluster Grouping; *Factor Analysis; Food Standards; *Measurement Techniques; Research Methodology; Taxonomy; *Testing
IDENTIFIERS	*Yeast

ABSTRACT

This is a methodological study that suggests a taxometric technique for objective classification of yeasts. It makes use of the minres method of factor analysis and groups strains of yeast according to their factor profiles. The similarities are judged in the higher-dimensional space determined by the factor analysis, but otherwise rely on the simple concept of "most like" or "neighbor." The proposed techniques are illustrated by means of an example involving 110 strains of yeast with measurements on 30 variables. An analysis in terms of six factors is obtained and the six-dimensional factor profiles for the strains are the basis for determining neighbors and classifying the strains into groups. The automatic procedure leads to 32 groups. Then, by applying the procedures again, only eight second-order groups, or clusters, emerge. (Author)

RB-71-65

RESEARCH BULLETIN

HOW FACTOR ANALYSIS CAN BE USED IN CLASSIFICATION

Harry H. Harman

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
December 1971

How Factor Analysis Can Be Used in Classification

Abstract

This is a methodological study that suggests a taxometric technique for objective classification of yeasts. It makes use of the minres method of factor analysis and groups strains of yeast according to their factor profiles. The similarities are judged in the higher-dimensional space determined by the factor analysis, but otherwise rely on the simple concept of "most like" or "neighbor." The proposed techniques are illustrated by means of an example involving 110 strains of yeast with measurements on 30 variables. An analysis in terms of six factors is obtained and the six-dimensional factor profiles for the strains are the basis for determining neighbors and classifying the strains into groups. The automatic procedure leads to 32 groups. Then, by applying the procedures again, only eight second-order groups, or clusters, emerge.

How Factor Analysis Can Be Used in Classification[1]

Harry H. Harman

I certainly do not consider myself an expert in taxonomy -- much less in microbiology or yeasts -- but perhaps I can make a small contribution to the problem of classification of yeasts. Mathematics and statistics -- as tools dealing with abstractions rather than actual substance -- have such universality that often their development in one discipline can find ready adaptation and application to an entirely different discipline. It is in this sense that I hope my theoretical work in factor analysis and my experience in applying these methods in psychological and educational measurement will be equally beneficial in resolving some of the classification problems of concern to this Conference.

Even my brief exposure to this field has shown me that it has been well plowed by many experienced and devoted workers. Among the many endeavors, let me mention only a few that are somewhat related to my approach: Hill (1965), Kocková-Kratochvilová (1969), Lance and Williams (1967), Pokorná (1969), Sokal and Rohlf (1970), Sokal and Sneath (1963). In addition to the work being done in biology and taxonomy, per se, there is a vast literature in educational and psychological measurement on the subject of grouping criteria and methods, on similarity profiles, etc. A quick sampling of such works will illustrate the point (e.g., Cronbach and Gleser, 1953; Harris, 1955; Johnson, 1967; McQuitty, 1956; Ward, 1963).

My understanding of the general objective of taxonomy is to effect an orderly or scientific classification of certain entities (e.g., biological) according to their presumed natural relationships. More specifically, the task is to assign each entity or element to a group such that there is a well-defined basis for "belonging to a group" and that the groups are clearly distinguishable one from another. It is generally assumed that there are a very large number of elements in the original set and that the number of groups is small by comparison. Put simply, then, we seek a means for allocating each element to a group, in some objective sense, so that the grouping is the best possible.

[1] Paper presented at the International Symposium: Yeasts as Models in Science and Technics, Smolenice -- Castle near Bratislava, June 1-4, 1971.

What I intend to do is bring together the technique of factor analysis and the technique of similarity grouping to provide an objective means of classifying elements (yeasts, in particular). My approach will include some broad philosophical considerations, some specific mathematical methods, and an indication of the computer procedures available; it will be illustrated with empirical data on yeasts made available to me by Dr. Kocková-Kratochvilová. I make no presumptuous claims for what I propose -- its potential value in your field is for you to judge in due time.

THEORY AND METHODS

It should be noted from the outset that factor analysis is not presumed to yield fundamental, primary or ultimate, entities; rather, factor analysis is a technique that yields descriptive categories or classification schemes for a set of data. Furthermore, different schemata of classification may appropriately be made for the same data.

Factor analysis can be of much help to the investigator if he is trying to understand and describe the relationships among many variables (or characters). The emphasis here is on the multi-dimensional relationships. An investigator frequently works in a higher-dimensional space but draws conclusions from relationships between pairs only, because that can be visualized and handled simply.

To show how factor analysis can be used for classification purposes, let us start out by defining explicitly some of the basic concepts, as summarized in Table 1. The N sampling elements can be represented by N points in an n-space (a hyper-ellipsoid, corresponding to a scatter plot representing a correlation between two variables in a plane); or, alternatively, as n vectors in an N-space, where the correlation between any two variables is given by the cosine of the angle between them (Harman, 1967, pp. 61, 96-97). Whichever geometric representation is assumed, it certainly does not require more than the lesser number of dimensions to account completely for all the interrelationships of the variables. For practical purposes it can usually be accomplished with a very much smaller number of dimensions, or common factors (Harman, 1967, Theorem 4.6, p. 63).

TABLE 1

Basic Concepts of Factor Analysis and Classification

Concept		Order	Example	Description
Name	Symbol			
Individuals (strains)	i	N	110	Sampling elements or entities
Variables (characters)	X_j	n	30	Observed measures
Data matrix	X Z	$n \times N$	30×110	Observed data raw scores: X_{ji} standardized: z_{ji} with M=0,S.D.=1
Correlation matrix	R	$n \times n$	30×30	Relationships among variables
Factors	F_p	m	6	Theoretical constructs (latent variables)
Factor matrix	A	$n \times m$	30×6	Coefficients of m common factors
Factor scores	$\hat{F}_{pi}$	$m \times N$	6×110	Profile of each individual (strain) in terms of m factors; when no confusion, the hat is omitted
Groups	G_j	$<N$	32	Basic grouping of individuals (strains)
Clusters	C_k	much less N	8	Higher-order grouping of individuals (strains)

The basic model of factor analysis may be put in either algebraic or matrix form:

(1) $$z_j = a_{j1}F_1 + \dots + a_{jm}F_m + d_jU_j \quad \text{or} \quad Z = AF + DU$$

where d_jU_j or DU represent the unique (specific and error) portions of each variable and are of little concern to us, while the m common factors are involved in fitting the correlations among all the variables and are of primary concern. The factors (F's) are theoretical constructs arrived at indirectly through the known relationships (the correlations) among the observed variables. The immediate object in performing a factor analysis is to get the coefficients of the factors in (1), that is, the factor matrix A.

Before continuing with the analysis, I want to stress that when I speak of factor analysis, I mean factor analysis -- not component analysis.[2] All too often studies are reported in which the investigator obtains principal components because such a computer program was readily available, when he should have obtained (or thought he was obtaining) factors according to the model (1). Hopefully, by using the model (1), we eliminate extraneous variance (error and specific) that would muddy up the explanation of relationships among the characters, and the consequent grouping of strains.

I also want to say a few words about which correlations are used in factor analysis. In the area of numerical taxonomy, Sokal and Sneath (1963, p. 208) note that "...in work done so far usually fewer OTU's [operational taxonomic units] than characters have been measured. It has been simpler, because of limited capacity of computers, to calculate correlations among OTU's than correlation among characters. As computational equipment gets better and faster, we shall be able to attack these problems more efficiently." That time has arrived. We do have the necessary computational equipment, and there is no longer any reason for compromising the statistical methods. By working with the correlations among the characters we avoid problems of deficient rank that would arise in a matrix of correlations among the strains when these exceed the number of characters. More importantly, factor analysis of the characters requires knowledge of the characters for scientific interpretation; factor analysis of the OTU's (or strains) requires knowledge about these elements themselves. But a major purpose of classification is to be as objective as possible in allocating the elements to groups. The principle of objectivity seems to be served better by determining the factors from the relationships among the characters.

A general description of the analysis proposed may be put in the geometric terms introduced above. Assuming a reasonably good fit of the N points in the n-space by the m common-factor space, then each of the N points can be expressed in terms of m coordinates, that is, by an m-order vector. Of course, m is much smaller than n. This reduction -- describing the strains in terms of m factors instead of in terms of the n observed characters -- can be

[2] For further discussion of the distinction between the classical factor analysis model and the component analysis model see Harman (1967), pp. 14-16, 136-137, 346-348.

accomplished by conventional factor analysis of the n characters and getting factor measurements, or scores, for each of the N strains. Finally, the N strains are classified into groups according to their similarities as determined by their factor profiles in the m-space. An advantage of the use of factor analysis in this way is that the characters themselves are structured so that the groups into which the strains are classified can be given special interpretation.

We come to the fundamental question: how is the factor matrix A determined? There are several procedures (other than component analysis) that are suitable. I prefer the "minres method," which is designed to give the best fit to the observed correlations, or to give minimum residual errors. Specifically, the minres method determines A under the condition (Harman, 1967, p. 189);

$$(2) \qquad f(A) = \sum_{k=j+1}^{n} \sum_{j=1}^{n-1} (r_{jk} - \sum_{p=1}^{m} a_{jp}a_{kp})^2 = \text{minimum.}$$

It should be noted that this expression depends on the number of common factors m. All procedures for getting factor solutions require a priori choices of either the communalities or the number of common factors. While the minres method requires a decision on m (and the computer program permits several values to be tried), the communalities are obtained as a by-product of the method. The mathematical theory for minimizing the objective function (2) has been developed (Harman, 1967, pp. 190-199) as well as an efficient computer program for the calculation of A.

After the common-factor space has been determined by the minres method, it is usually advisable to select another frame of reference for purposes of interpretation. The varimax method (Harman, 1967, pp. 304-313) can serve that purpose. At this stage, the structure of the characters can be used to provide meaning for the factors which emerge as theoretical constructs. Although not directly measurable, the factors scores can be estimated (Harman, 1967, pp. 350-354) for each individual or element. These profiles of factor scores serve as the basis for judging similarities among the elements.

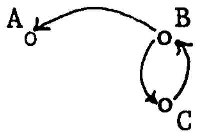
Then the actual task of classification according to this basis must be performed. In order to group elements (e.g., strains of yeast) according to their similarities, there immediately arises the question of the precise meaning of "similarity." It would seem natural to accept two elements as similar if they resembled one another or were close to one another in some sense. When the elements are described in terms of some quantitative characteristics the natural approach is to compare each element with every other one and to say that those with profiles closest to one another are "similar." A measure of closeness frequently employed by researchers is that of "distance." Thus, if two elements are represented by two points in the m-dimensional space, the square of their Euclidean distance is simply the sum of squares of the differences between corresponding numbers in their profiles. Then small distances can be used as a measure of similarity, while large distances would indicate dissimilarity. But care must be taken that all variables are measured in essentially the same scale; variables measured on large scales (i.e., with large standard deviations) could influence distance measurements unduly. This is avoided when factor scores are used since they are essentially equivalent scales.

The classification procedure that I will employ is due to Wingersky (1969) and rests on the basic premise that a given individual and the individual most like him should be classified in the same group. For the similarity basis it is easier to define the complement, or <u>dissimilarity</u>, as given by the squared Euclidean distance:

$$(3) \qquad D_{ij} = \sum_{p=1}^{m} (F_{pi} - F_{pj})^2 \qquad i, j = 1, 2, \ldots, N$$

where the hat has been left off the symbol for the factor scores of elements i and j . Of course, the smaller this value the more similar are the two elements. While the distance itself (instead of its square) might be used, I was willing to have the dissimilarities appear exaggerated in order to show the logical cost of grouping.

It is important to note that the "most like" relationship between two elements is <u>not</u> reciprocal -- the individual most like A may be B but this does not necessarily mean that the individual most like B is A. To illustrate this point, suppose three towns A, B, and C are situated so that B is 10 km. east of A and C is 5 km. south of B, while no other town is as close as 10 km. to any of these. Now, using the term "nearest" or "neighbor" in place of "most like," we may say that B is nearest or is the neighbor of A, but A is not the neighbor of B; C is the neighbor of B and B is the neighbor of C. If we were to classify towns on the basis of their similarity -- a town and its neighbor should be in the same group -- then A and B would be placed in the same group and B and C would be in the same group, and hence A, B, and C must be in the same group. We shall represent the relationship "B is the neighbor of A" by B → A. Hence, the illustration of the three towns may be represented in the sketch.



Using our definition of similarity (i.e., the smallness of D_{ij}), a list of all elements can be formed showing the neighbor for each one. Then the classification procedure, which has been programmed by Wingersky (1970), can be applied. In essence, it works like this: the first element and its neighbor are taken to start the first group; additional elements are added to this group by scanning the list for elements that have neighbors or are neighbors of elements already in the group; the scanning of the list is repeated until no new elements can be added in accordance with the foregoing rule. After one group is closed, the remaining elements are treated as a new sample with one element and its neighbor selected to initiate a new group, additional elements are added by scanning the remaining list until a second group is formed. This process is continued until all elements have been assigned to groups.

RESULTS

It should be clear that my presentation is methodological rather than substantive. Nonetheless, I want to illustrate the methods with an example in as much detail as limited space will allow. The example is taken from some work in which Dr. Kocková-Kratochvilová and I are collaborating. That work will appear as a report on 110 strains of genus Säccharomyces, in which the detailed statistical procedures and results will be given. I had no previous classificatory knowledge about these strains in arriving at an entirely objective grouping.

The data seemed to be meaningful and therefore worthy of statistical analysis. As noted by Sokal and Rohlf (1970, p. 316): "When one performs a factor analysis of correlations of characters within a single homogeneous population, factors may represent various physiological and growth trends found within the population." The suitability of our data was indicated in a letter from Dr. Kocková-Kratochvilová, "... in the case of Säccharomyces, where the species are very relative and the genus seems to be homogeneous ...." Her further assurance that she "... selected the taxonomic characters very carefully" made the objective analysis reasonable.

Thirty characters were measured for each of the 110 strains. Space restrictions will not permit the presentation of the 30 x 110 data matrix, nor the 30 x 30 correlation matrix, nor the 30 x 6 minres factor matrix. The final 30 x 6 varimax factor matrix is presented in simplified fashion in Table 2. Looking down one column at a time, it should be possible to assign a name or describe each of the factors on the basis of the very high positive (+ +) and very high negative (--) weights; the smaller weights (+ or -) should fit in consistently. The blank entries represent insignificant weights and probably should not influence the description of a factor, but again should be consistent with it. The interpretation of these factors is left for our later work.

The structure of the characters can be inferred from this table. Aside from identifying those characters primarily responsible for the makeup of the factors, certain statistical properties about the characters themselves become apparent. Some of the characters are factorially simple (e.g., 3, 6, 13, 17, 30) while others are complex (e.g., 4, 8, 20). Characters that have very little in common with others in the study, such as 10 and 11 (indicated by their low communalities), might be eliminated in further investigations aimed at gaining a better understanding of these strains of yeast through statistical analysis.

The next step in the analysis produces a factor-score profile for each strain (again, space does not permit the presentation of this 6 x 110 matrix). Also calculated at this time are the regression equations which yield these profiles and the multiple correlation of each factor as predicted from the 30 variables. Once the factor profiles are available, the squared distance between

TABLE 2

Prominent Weights on Six Varimax Factors
(Initial solution: Minres)

	Character	F_1	F_2	F_3	F_4	F_5	F_6	Communality
1	Mean of lengths of cells	++						.71
2	Mean of widths of cells	++					-	.86
3	Quotient surface/volume of cells	--						.83
4	Degree of raffinose fermentation	+			-	+	+	.65
5	Galactose fermentation					++		.52
6	Fermentation types (maltose and sucrose)	++						.62
7	Growth at 42 °C				+			.27
8	Osmophily	-	+	+	+			.50
9	Autoproteolytical activity	+	-					.37
10	Giant colony character							.08
11	Radial growth rate at 20 °C after 7 days							.05
12	Pseudomycelium formation					+	+	.35
13	Trehalose assimilation		++					.70
14	Inulin assimilation		+					.23
15	Mannitol assimilation		++					.62
16	Sporulation activity	+						.27
17	Requirement of vitamins						++	.49
18	Sensitivity to lactic acid		-		+			.58
19	Lactic acid dehydrogenase activity			+				.40
20	Sensitivity to actidione	+			+	-	+	.49
21	Sedimentation rate of cells				---			.42
22	Tolerancy to ethanol			+	+			.44
23	Galactose respiration quotient, RQ					+		.29
24	Glycerol assimilation						+	.20
25	Maltose utilization rate	+			+			.45
26	Sucrose utilization rate					+		.27
27	Agglutination with the serum against Saccharomyces cerevisiae			+				.32
28	Lysin assimilation	--						.52
29	Catalase activity	--		+				.55
30	Succinic acid dehydrogenase activity			++				.62
	Contribution of factor	4.70	2.28	1.97	1.82	1.46	1.44	13.67

Key: ++ $a \geq .60$
+ $.60 > a \geq .30$
- $-.30 \geq a > -.60$
-- $a \leq -.60$

each strain and every other one is readily computed; a list of all the strains, showing the neighbor of each, is formed; and the classification procedure groups the strains. The results are still too lengthy to include here, but one subset should help to clarify the nature of the analysis.

In Table 3 are exhibited twelve strains, with the neighbor of each, and four groups into which they are classified. The six-factor-score profile is shown for each strain and for the centroid of each group. The centroid profile may be considered as representing the group. (The last line of the table will be explained shortly.) Also included in the table are (1) the squared distances between every pair of strains within each group, (2) the squared distance between each strain and the centroid of its group, and (3) a measure of cohesiveness of the grouping. For the latter measure it is easier to define the complement, or <u>separateness</u>, simply by the average of all the squared distances among all pairs within a group G_j, namely:

$$(4) \qquad S_j = \sum_{p<q} D_{pq}/\nu_j \qquad \text{and} \qquad \nu_j = \frac{1}{2} n_j (n_j - 1) ,$$

where p and q range over the elements in G_j and n_j is the number of elements in this group. Here again, the smaller the value of S_j the more cohesive are the elements in the group.

The information of Table 3 may be shown graphically only in rough fashion, as in Figure 1. The true relationships of the analysis are in a six-space and so we can't visualize it in the ordinary way. The actual squared distances (D_{ij}) between the strains within each group are given in Table 3 and the squared distances between the group centroids are shown in the little table in the figure. What appears in Figure 1 is a projection of the six-space configuration on a plane. Only the first two coordinates of each six-component vector is used. Therefore, the squared distances (D_{ij}) actually used in the analysis will not agree with corresponding measures in the plane. It is all right to use such graphs for general impressions, but not to draw precise inferences.

TABLE 3

Classification of Strains within Groups of Cluster 2, including Squared Distances

Strain	Neighbor	Factor Profiles						Squared Distances within Groups			
		F_1	F_2	F_3	F_4	F_5	F_6				
						Group 2					
2:(21-2-1)	69	-.86	.03	-.14	-.27	-1.88	.23			S_2 = .84	
4:(21-2-3)	5	-1.25	-.09	.02	-.67	-1.68	-.92	1.70			
5:(21-2-4)	4	-1.01	-.12	.07	-.50	-1.68	-.75	1.14	.12		
69:(35-6-2)	5	-1.24	-.35	-.42	-.74	-1.81	-.34	.91	.61	.59	
Group Centroid		-1.09	-.13	-.12	-.55	-1.76	-.45	.62	.29	.14	.21
						Group 3					
3:(21-2-2)	74	-1.70	.95	-.37	-.67	-.63	-1.03			S_3 = 1.88	
32:(21-23-1)	3	-1.82	.38	-1.05	-.30	-1.07	-.23	1.76			
74:(35-9-1)	3	-.95	.44	.09	-.69	-.55	-.67	1.20	2.69		
Group Centroid		-1.49	.59	-.44	-.55	-.75	-.64	.36	.86	.67	
						Group 8					
26:(21-22-1)	30	-2.47	-.36	-.71	-.05	.82	-1.65			S_8 = 1.43	
29:(21-22-4)	30	-2.50	-.87	-1.61	.78	.73	-1.18	2.00			
30:(21-22-5)	26	-2.63	-.42	-1.53	.30	.76	-2.08	1.01	1.27		
Group Centroid		-2.53	-.55	-1.28	.35	.77	-1.63	.53	.62	.28	
						Group 9					
27:(21-22-2)	28	-1.81	-.78	-.91	-.01	-.23	-.98			S_9 = .70	
28:(21-22-3)	27	-2.00	-.50	-.37	-.15	-.25	-.46	.70			
Group Centroid		-1.91	-.64	-.64	-.08	-.24	-.72	.17	.17		
Cluster 2 Centroid		-1.69	-.14	-.58	-.25	-.62	-.84				

Squared Distances

Grp	2	3	8	9
2			S_{C_2} = 5.02	
3	1.85			
8	12.22	7.18		
9	3.81	2.21	2.84	
C_2	2.11	.74	4.30	.49

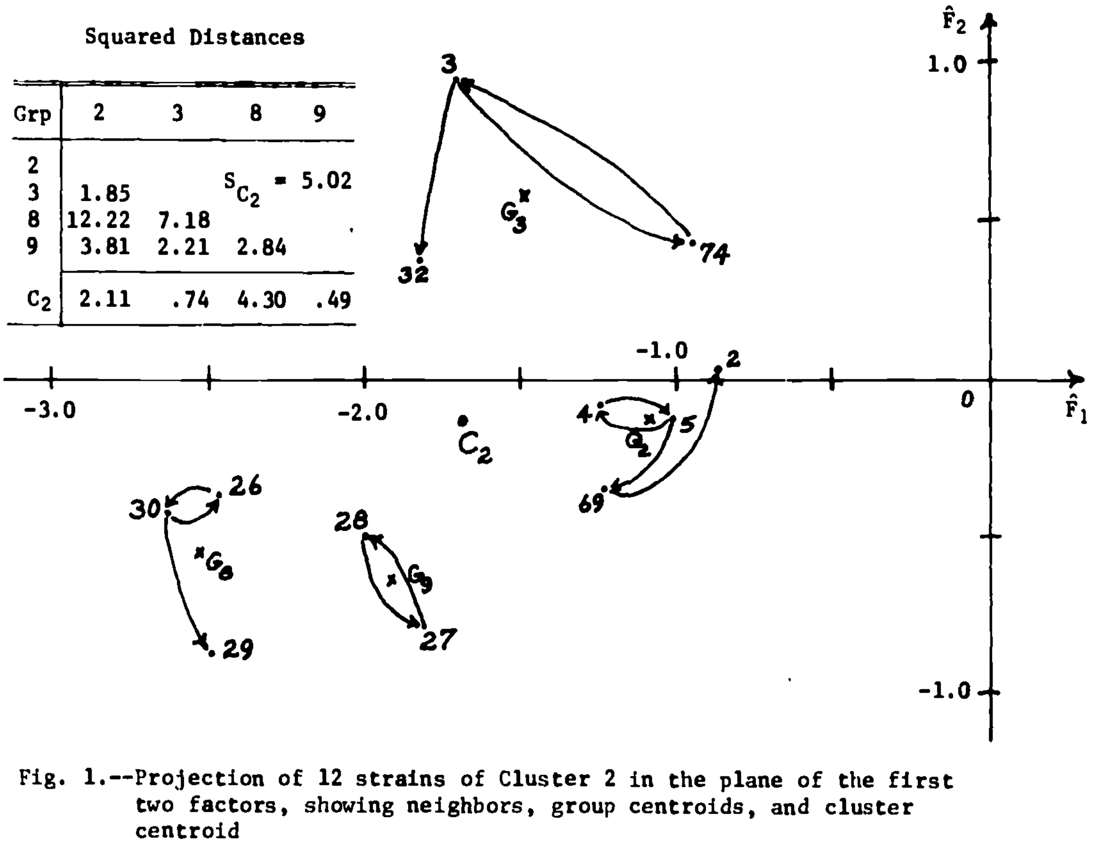


Fig. 1.--Projection of 12 strains of Cluster 2 in the plane of the first two factors, showing neighbors, group centroids, and cluster centroid

For the total sample of 110 strains 32 groups resulted, which are shown with their (centroid) profiles in Table 4. The classification procedure tends to produce too many groups. If one is willing to make a compromise -- some loss in cohesiveness in order to gain simplification in the descriptive model -- the method can be applied again. This time, to distances between (centroids of) groups instead of to distances between strains. The results of this "higher-order" classification are being called "clusters."

When the classification procedure was applied to the 32 groups as the sample elements, eight clusters emerged as shown in Table 5. Following the factor profiles are the numbers of the groups and their neighbors (by the arrow convention introduced above). The actual numbers of the strains can be read from Table 4 for the groups in each cluster. Finally, the squared distances between each pair of clusters is shown in the right-hand part of Table 5.

Earlier, the meaning of the last line of Table 3 was postponed. Now that we have defined second-order classification, we can understand the example of Table 3, which gives such details of Cluster 2 as the factor score profile for each of its 12 strains and for each of its four groups. The factor profile for the entire Cluster 2, shown in the last line is, of course, the same as that shown on the second line of Table 5.

We might take a moment to look at Cluster 2 in relation to the other clusters in Table 5. Its neighbor (closest cluster) is C_4 with $D_{C_2C_4} = 4.50$, and the farthest cluster is C_7 with $D_{C_2C_7} = 14.02$.

The measure of separateness, formula (4), can be applied to the elements (i.e., groups) of a cluster. Thus, for Cluster 2 the degree of cohesiveness is given by the value 5.02 (shown in the table in Figure 1). Of course, this value is larger than the S_j values for the groups in this cluster. When the measure of separateness is computed for the clusters as elements, the value S = 8.14 is obtained -- an exceedingly high value as we should expect. This is a kind of "third-order" classification of all eight clusters into a single family.

A wealth of information is summarized compactly in Tables 2 through 5. Space limitations precluded any elaboration of the results. Careful study by those concerned with statistical methods of classification may find the methods and means of display quite rewarding.

TABLE 4

Classification of 110 Strains into 32 Groups,
Showing Factor Profiles for Group Centroids

Group	No. of Strains	Factor Profile						Strains in Group
		F_1	F_2	F_3	F_4	F_5	F_6	
1	9	-.12	-.01	1.03	-.21	-.13	.48	1,7,12,19,20,21,23,36,37
2	4	-1.09	-.13	-.12	-.55	-1.76	-.45	2,4,5,69
3	3	-1.49	.59	-.44	-.55	-.75	-.64	3,32,74
4	4	-1.22	-.13	.65	.03	1.07	-.20	6,14,25,64
5	2	-.18	-.26	.06	-.05	.09	1.90	8,9
6	3	-.75	-.58	1.33	.44	.82	.65	10,15,16
7	6	-.45	-.32	1.79	-.21	.41	.46	11,13,17,18,22,24
8	3	-2.53	-.55	-1.28	.35	.77	-1.63	26,29,30
9	2	-1.91	-.64	-.64	-.08	-.24	-.72	27,28
10	2	.20	.08	.83	-.23	.40	-.78	31,54
11	3	-.37	-.14	-.37	.84	-.72	-.04	33,41,44
12	2	-1.45	.34	.10	-1.69	1.07	-1.14	34,55
13	3	.55	.10	.21	.61	-.11	-.25	35,48,49
14	4	-.29	3.29	-.32	.12	.41	-.32	38,51,52,65
15	4	.44	.09	-.42	.36	.56	1.54	39,40,42,53
16	3	.04	-.48	-.61	.09	-1.53	.06	43,45,46
17	3	.17	-.32	-.79	.31	-.35	.57	47,50,72
18	3	.05	2.89	-.52	.02	-.95	.92	56,70,73
19	5	1.04	-.18	.84	.39	-1.23	-.92	57,59,92,94,95
20	2	.18	1.33	-.41	.07	-.05	.68	58,71
21	5	.10	-.52	-1.12	-.97	.99	.61	60,75,76,78,79
22	3	.53	-.30	.12	-.21	-.39	.84	61,62,66
23	4	-.14	-.37	.36	.74	.62	.56	63,104,105,106
24	2	-1.23	-.85	-.79	.54	-1.69	.65	67,68
25	2	.84	-.92	-1.35	-.86	.19	1.68	77,82
26	2	.74	-.81	-1.19	-1.26	.68	1.20	80,81
27	4	1.37	-.28	-.34	-2.26	.26	-.95	83,85,86,88
28	2	1.65	-.04	-.91	-1.74	-.05	-1.65	84,87
29	2	1.05	-.46	-.03	.89	-1.16	-.61	89,91
30	2	.34	-.46	.32	.68	-1.14	-.59	90,93
31	6	.90	-.34	-.46	.97	.49	-.71	96,98,99,101,102,103
32	6	1.23	-.12	.01	1.40	.97	-.79	97,100,107,108,109,110

TABLE 5

Classification of 110 Strains into Eight Clusters via 32 Groups, and Squared Distances among Clusters

C_k	Factor Profile						Groups in Cluster	No. of Strns	Squared Distances among Clusters							
	F_1	F_2	F_3	F_4	F_5	F_6			1	2	3	4	5	6	7	8
1	-.51	-.17	1.01	-.08	.45	.29	1,4,6,7,12,23,	28								
2	-1.69	-.14	-.58	-.25	-.62	-.84	2,3,8,9	12	6.34							
3	.05	1.62	-.35	.14	.07	.85	5,14,15,18,20	15	5.88	9.67						
4	-.10	-.39	-.47	.30	-.88	.40	11,16,17,22,24	14	4.33	4.50	5.23					
5	.87	-.14	-.03	.91	.54	-.66	10,13,31,32	17	4.90	9.59	7.00	4.74				
6	.41	-.67	-1.19	-1.01	.74	.98	21,25,26	9	7.35	10.80	7.88	5.56	8.31			
7	1.46	-.20	-.53	-2.09	.16	-1.18	27,28	6	12.56	14.02	14.45	11.74	10.04	7.94		
8	.89	-.30	.53	.56	-1.19	-.78	19,29,30	9	6.47	8.87	9.63	3.52	3.49	12.64	10.48	

S = 8.14

DISCUSSION

In this paper we stressed the necessity for distinguishing between (1) the basis for judging similarity of elements and (2) their actual classification according to a designated basis. The basis proposed was the factor profiles and the classification procedure required that a given element and another most like it be classified in the same group. The very concept of grouping arises out of the scientific aim of deriving underlying orderliness in otherwise diverse observations. Recognizing that observed data are fallible, we believe simple models provide more appropriate explanations than accepting every fine difference as signifying something "real." Thus, we seek a simpler explanation for all the distinctions among the 30 separate measurements of the 110 strains (3300 original observations). The analysis first reduced the 30 characters to six hypothetical constructs (leading to 110 x 6 = 660 factor measurements which account for 45% of the total variance). Then the 110 strains were put into 32 groups and finally into eight clusters (8 x 6 = 48 or about 1.5% of the original number of categories). Of course, at each stage a simpler model is introduced, which might be viewed as "smoothing" of the data.

The classification of yeast strains into groups and clusters forms a hierarchical arrangement and might therefore be displayed as a dendogram. However, such a representation in the plane cannot show the distances among the elements when the basis for classification consists of six-dimensional profiles. Even in an excellent study (Pokorná, 1969) in which yeast strains were analyzed in terms of five common factors, the attempt to group the strains according to their clustering in a plane of two factors at a time could be misleading.

Finally, may I say, the classification of yeasts into groups and clusters, as demonstrated in this paper, was done more "objectively" than I would ordinarily recommend. I was 6000 kilometers from Dr. Kocková-Kratochvilová and could not have the benefit of frequent consultation about the meaningfulness of the classifications. It is perfectly good science -- I would say, imperative -- that the classifications be inspected for substantive sense, and adjustments made accordingly. Science should not be blind.

REFERENCES

Cronbach, L. J., and G. C. Gleser. Assessing similarity between profiles. Psychol. Bull., 50 (1953), 456-473.

Harman, H. H. Modern Factor Analysis. (2nd ed.) University of Chicago Press, 1967.

Harris, C. W. Characteristics of two measures of profile similarity. Psychometrika, 20 (1955), 289-297.

Hill, L. R., L. G. Silvestri, P. Ihm, G. Farchi, and P. Lanciani. Automatic classification of Staphylococci by principal-component analysis and a gradient method. J. Bacteriology, 89 (1965), 1393-1401.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 32 (1967), 241-254.

Kocková-Kratochvilová, A., J. Sandula, A. Vojtkova-Lepsikova, and M. Kasmanova. Taxometric study of the genus Saccharomyces (Meyen) Reess. First Part. Bratislava: Slovak Academy of Sciences, 1969.

Lance, G. N., and W. T. Williams. A general theory of classificatory sorting strategies. Computer Jrnl., 9 (1967), 373-380.

McQuitty, L. L. Agreement analysis: classifying persons by predominant patterns of responses. Brit. J. Stat. Psychol., IX (1956), 5-16.

Pokorná, M. Statistical analysis of micromorphological dimensions and inhibition of growth by antibiotics and acids in pathogenic species of the genus Candida. Folia Microbiologica, 14 (1969), 544-553.

Sokal, R. R., and F. J. Rohlf. The intelligent ignoramus, an experiment in numerical taxonomy. Taxon, 19 (1970), 305-488.

Sokal, R. R., and P. H. A. Sneath. Principles of Numerical Taxonomy. W. H. Freeman and Co., San Francisco and London, 1963.

Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assoc., 58 (1963), 236-244.

Wingersky, B. G. The classification problem. Unpublished paper presented at Western Psychological Association, Vancouver, Canada, 1969.

Wingersky, B. G. NABORS: A program for grouping profiles. Princeton, N. J.: Educational Testing Service, 1970.