Hazards in Factor Analysis

C. GARDNER SWAIN,* HENRY E. BRYNDZA, and MARGUERITE S. SWAIN

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received August 28, 1978

Principal components and other standard factor analyses can yield misleading results if an assumed subsidiary condition is untrue or if data are missing. This is illustrated by absurd results obtained from a problem with known answers. Even when test data are complete (e.g., 70 data on seven properties, dependent on both radii and heights of ten cylinders), principal components followed by varimax or other standard rotations gives incorrect rank orders for factors (factor scores for radius and height for each of the ten cylinders) and sensitivities (factor loadings for each of the seven properties). When no data are missing, a transformation incorporating valid subsidiary conditions can be used instead of such rotations to obtain correct factors and sensitivities, although no such transformations have been used in any previously published work. However, when a moderate number (e.g., 20, or 29%) of the possible data are missing (randomly deleted), factors and sensitivities can have wrong rank orders and therefore be misleading even with this transformation. When data are missing, standard factor analysis is evidently unreliable and should be replaced by another method, such as that in the preceding paper.

Factor analysis is now widely used to identify and evaluate significant underlying factors from masses of numerical data.[1] The methods were developed mainly by psychologists. By 1966 more than 2000 publications using factor analysis had appeared.[3] Since then, applications have mushroomed into such fields as psychiatry,[4] medicine,[5] anthropology and biology,[6] chemistry,[7] geology,[8] and national and international relations including military strategy.[9] Textbooks on factor analysis abound.[10,11] Applications are facilitated by the ready availability of computer programs for factor analysis in SPSS,[12] BMD and BMDP,[13] DATA-TEXT,[14] P-STAT,[15] SSP,[16] IMSL,[17] TSP,[18] and other software packages for statistical analysis.

Factor analysis was designed to find the underlying determiners (factors) that can account for the correlations between the different sorts (series) of data measured. Recently we became concerned about whether or not the standard factor analysis procedures do, in fact, succeed in this goal and, if not, what modifications are needed to make them succeed. It is difficult or impossible to answer these questions from examination of past applications because the true underlying factors are generally not known in real problems. Thustone's tests[19,20] were not adequate, as we shall show below. Therefore, we devised the following synthetic problem where a correct set of underlying factors can be distinguished from an incorrect set.

We attempted to analyze data on seven properties of right circular cylinders (areas, masses, moments of inertia, etc.), all of which are determined precisely by only two underlying types of factors (radii r and heights h). To be successful, a factor analysis must separate these factors. By analysis of these data, it must calculate a number (factor) for each cylinder that is a pure measure of cylinder radius. This may be r, $2\pi r$, $4\pi r^2$, $\log r$, $2 \log r$ or $k \log r$, but must not be rh, r/h or any other mixture or hybrid of r and h. It must calculate a second factor for each cylinder that is a pure measure of cylinder height. This may be h, $\log h$, or any other fixed function of h that is unaffected by r. It must not calculate or indicate any additional factors. Success or failure in this simple synthetic example is therefore clearly defined, and can be used to test the standard procedures commonly employed and various modifications of them. We shall show that the standard procedures fail to meet this challenge, but that success can be attained by adding a novel kind of transformation (rotation), provided that there are no missing data.

It has been recognized that the principal problem with the use of factor analysis lies in discovering an appropriate transformation (rotation) capable of separating the factors into pure (unhybridized) types. Rotation of axes and variation of the angle between axes have been tried in various combinations, but they have not been logically derived from the specific problem and do not give correct parameters. We shall show that *the transformation equations must be derived from valid (true) subsidiary information related to the problem at hand.* Unfortunately, no one has derived transformation equations from such problem-related subsidiary conditions previously. Instead, standard transformations such as varimax rotations have generally been used. Since we shall show that these fail to separate the factors in our simple example, we believe that they may have failed to separate them in all previous applications.

It has been recognized that missing data are undesirable. Gaps in the input data matrix introduce serious errors into the correlation coefficient matrix, well in advance of the factor extraction and transformation procedures. Unfortunately, many "factor analyses" have been carried out in spite of missing data. SPSS provides three different options for coping with missing data. We shall show that *factors are not separated correctly when data are missing*, using any of the standard methods for handling missing data, even with our novel transformation which works correctly with a full data set.

Our study is presented under four subheadings: (1) selection of a problem suitable for testing factor analysis, one where success or failure can be recognized but otherwise similar to a real application; (2) the transformation needed after the initial extraction process; (3) a demonstration that extraction followed by this transformation yields correct parameters when no data are missing; (4) a demonstration that seriously erroneous and misleading parameters can result if a moderate fraction of the pertinent data are missing (e.g., owing to their never having been measured, or to lost records, or incomplete returns, or deliberate omission due to concern about accuracy) even if all of the used data are perfectly accurate.

METHOD

Choice of a Test Problem. The hazards in factor analysis first became apparent to us when we attempted to apply principal components[1] to a chemical problem, to identify and evaluate the two solvent characteristics most responsible for

0095-2338/79/1619-0019$01.00/0 © 1979 American Chemical Society

Table I. Second-Mode Slopes, b_i, from Principal Components Analyses

						50 true + 20 calcd data			
	70 true data			50 true data		after transformation; no. of recalcns of 20 data from latest parameters			
i	not rotated	varimax rotated	after transformation	not rotated	after transformation	0	3	4	5
1	−0.04	0.56	(0.00)	0.03	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
2	0.22	0.33	1.00	0.26	0.78	0.93	0.88	0.88	0.90
3	0.38	0.16	(1.00)	0.47	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
4	0.11	0.43	1.00	0.02	−0.11	0.49	1.01	1.21	1.39
5	−0.74	0.98	−1.00	−0.80	−0.67	−0.59	−0.79	−0.74	−0.67
6	0.22	0.33	1.00	0.13	0.36	0.25	1.28	1.39	1.46
7	0.40	−0.82	1.00	0.42	1.18	0.85	1.06	1.11	1.08
%[a]	100.0	100.0	100.0	101.7	101.7	82.9	98.2	98.7	98.7
cc[b]	1.000	1.000	1.000	0.833	0.833	0.813	0.926	0.925	0.917

[a] Percent of the variance explained by two modes. [b] Correlation coefficient based on only the true data used.

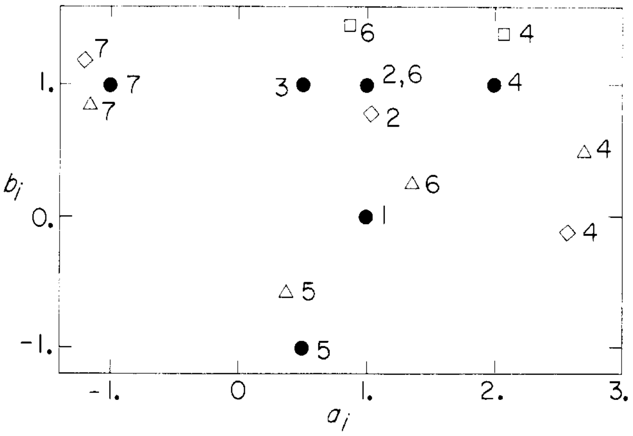

Figure 1. Solid circles ● show correct relative sensitivity to log height vs. relative sensitivity to log radius for seven cylinder properties, calculated from 70 data by principal components followed by a valid transformation. Also shown are the most seriously incorrect values calculated similarly except from 50 true data (with 20 data or 29% of the possible data missing) via pairwise deletion ◇, replacement by means △, or five subsequent iterative replacements □. Other calculated points falling closer to the correct values have been omitted.

changes in thermodynamic, kinetic, and spectral data when the solvent is varied. Missing data (never measured) were extensive in this problem, comprising 75% of the possible combinations of the 25 reactions and 60 solvents considered, but the data that were available covered ranges of more than a power of ten for each reaction and were believed to be individually accurate and precise to about ±15%. Principal components gave solvent parameter values in absurd rank orders using any of the programmed procedures for handling missing data. Therefore, we undertook to examine the method further, using a similar but simpler problem. Since logarithms of the measurements constituting the data in our chemical problem appeared to be linear functions of two solvent factors (anion-stabilizing ability and cation-stabilizing ability), we chose a test problem where logarithms of the data were known to be accurately linear functions of two factors, i.e., the cylinder problem that was solved correctly, without or with missing data, by our "DOVE" (dual obligate vector evaluation) procedure outlined in the previous paper.[2]

Thurstone attempted to justify factor analysis by means of a cylinder problem more than 30 years ago,[19] in one of the few recorded tests[20] that involved applying it to a problem with answers known in advance, but neither his cylinder problem nor any of the other test problems had any missing data. Thurstone's problem was based on diameters and heights carefully preselected to have zero correlation in spite of small sample size. Real data samples seldom have this property. Nevertheless, his selection of an application to cylinder properties was an inspired choice for testing factor analyses because it embodies the most essential features of a real two-mode problem in an especially simple and easily understood form.

The Transformation. Slopes (a_i,b_i) and factors (x_j,y_j) obtained initially from principal component analysis with two modes $(n = 2)$ correspond to formulation of the data as $p_{ij} = a_ix_j + b_iy_j$ under the conditions that the factors are orthogonal, i.e., $\sum_j x_jy_j = 0$ for all j for which data exist (e.g., for our ten cylinders), and that the slopes are normalized, i.e., $a_i^2 + b_i^2 = 1$ for each i (property). However, this orthogonality condition is undesirable, and a condition of zero covariance or zero correlation coefficient between the factors would also be undesirable, because none of these is likely to be even approximately true for a small sample (only ten cases). Furthermore, in many other problems none of these would be a good approximation even for thousands or even an infinite number of cases. For example, in our solvent effect problem a weak negative correlation between the two factors is expected and perfectly reasonable. The second condition, $a_i^2 + b_i^2 = 1$, is also generally untrue. Therefore we should replace these two untrue conditions by two valid conditions.

Our transformation is carried out as follows. The old a_i and b_i values from factor analysis (SPSS,[12] unrotated) are unstandardized to make them comparable with the observed data, z_{ij}, by multiplication of each by the standard deviation of z_{ij} for its series (i) from the mean of z_{ij} for that i. Corresponding x_j and y_j values are calculated by least squares from the original data z_{ij} and the newly unstandardized a_i and b_i. Values of c_i are calculated by simple least squares to fit the observed data using $p_{ij} = a_ix_j + b_iy_j + c_i$. To obtain unique parameters, six independent subsidiary conditions must be specified.[2] We choose $x_5 = 0$, $y_5 = 0$, $a_1 = 1$ and $b_3 = 1$ as the four relatively unimportant ones (which define reference points and unit sizes). For the two critical statements, we again choose $b_1 = 0$ and $a_3 = a_1/2$ for logical reasons explained previously.[2] New parameters meeting the new conditions are calculated by the transformation equations (10–15) using t values derived previously.[2]

This transformation of the factors and slopes to meet the new six subsidiary conditions is necessarily considerably more complicated than the usual "rotations" of factor analysis which only rotate the axes in the plane on which the slopes are displayed.

RESULTS

With No Missing Data. Table I and Figure 1 show selected parameters after various procedures. We have plotted slopes here rather than factors because slopes (factor loadings) are more commonly reported by users of factor analysis than the factors (factor scores) themselves. From a complete data set, correct slopes are obtained after our transformation (column

4), although the values before transformation (column 2) or after standard rotations seem quite different. Values in parentheses are those assigned by the subsidiary conditions. Use of $a_7 = -a_1$ instead of $a_3 = a_1/2$ as the sixth condition yields identical values for all parameters. Introduction of small (~2%) random errors into the 70 data causes only small (~2%) errors in the factors and slopes.

Since most of the thousands of publications applying factor analysis have certainly not used our complicated transformation or its equivalent, but only varimax or other similar kinds of rotations contained in statistical analysis computer packages,[12-18] it is instructive to note what kind of results such standard factor analyses give.

Assumptions built into the principal components direct solution cause the distribution of data in space to determine the order in which modes are extracted and force the factors to be orthogonal and the sum of squared slopes to be unity, although these conditions may not correspond at all with reality. Other extraction or direct solution methods assume other conditions, similarly arbitrary or untrue. The very popular varimax rotation technique then "simplifies" the distribution of slopes so as to concentrate each slope on a few i's and in one mode, enlarging slopes for those i's and reducing them for others in that mode, again regardless of whether this makes physical sense or not. Other rotation techniques enhance this compartmentalized maldistribution of slopes even more.

Columns 2–4 show that 100% of the variance is explained (86.2% by the first mode, 13.8% by the second). Thus, principal components calculates correctly that there are two modes. However, this has no bearing on the correctness of the unrotated or rotated slopes.

Column 3 shows the results of varimax rotation, because this is the rotation method used in most published applications.[21] The second-mode slopes b_i are positive for $i = 2$, 4, and 6, as they must be if they are to make any physical sense, because the values of these properties do increase with increasing cylinder height. However, they certainly appear wrong for $i = 1$ and 7. The sensitivity of the first property, logarithm of the total area of the flat faces, to the second factor, logarithm of cylinder height, is deduced to be also positive and relatively large (0.56), *although we happen to know that it is, in fact, exactly zero.* The sensitivity of the last property, the log of electrical resistance between the ends, to log h is deduced to be even larger, near-unity, but *negative* (−0.82). Accordingly, electrical resistance of wires might be expected to decrease with increasing length, with a nearly reciprocal dependence. This would be a real boon for long-range power transmission and could lead to some interesting science fiction, but electrical engineers should not be misled by these results. Neither should those who would save energy be misled, by a thermal resistance analogy, into using close-packed lateral arrays of wide but very short cylinders (disks) for space-saving thermal roof or wall insulation in building construction.

Experimental verification of hypotheses is sufficiently practicable in physics and engineering that there has been little desire for and use of factor analysis in those fields. The greatest dangers are therefore in psychology, psychiatry, medicine, and political science, where experimental design is more difficult and where so many conclusions have accordingly been drawn from just such standard factor analyses with conventional rotations. It seems distinctly possible that progress in those fields has been hampered rather than helped in the past by such numerology, because *such misleading results are obtained even when no data are missing.* Such answers can be worse than no answers at all. Our test might be criticized because of the small number of types of factors in our example ($n = 2$, only two modes of interaction), whereas the problems to which factor analysis are applied usually have n equal to 3 or more. However, the untrue conditions corresponding to these artificial manipulations of the slopes, which are so obviously invalid in our example, are still assumed when there are more factors, although the absurdity of the results is then obscured by their greater complexity.

To obtain meaningful slopes (column 4) from factor analysis we have had to resort to our "transformation", instead of any conventional kind of rotation, in order to replace the untrue conditions by true ones, even when there are no missing data. Unfortunately, in other studies where factor analyses have been used, untrue conditions have generally not been replaced by a subsequent transformation incorporating true subsidiary conditions. Results of such studies are therefore unreliable except under special circumstances.[22]

With 20 Missing Data. Parameters deduced from 50 (accurate) data instead of 70, also with our tranformation using $a_3 = a_1/2$ as the sixth subsidiary condition, are shown in columns 6–10 of Table I. The 20 data that are missing (randomly deleted) are indicated in Table II.[2] The available factor analysis programs such as those in SPSS[12] provide options for the handling of missing data: listwise deletion, replacement by zeros, or pairwise deletion. Since listwise deletion eliminated all j except $j = 3$, and replacement by zeros gave worse results than pairwise deletion, columns 5–6 were obtained via the third option. The correlation coefficient between the 50 observed and calculated data is now only 0.833. The rank order for b_i is now $5 < 4 < 1 < 6 < 2 < 3 < 7$ instead of the true order, $5 < 1 < 2 = 3 = 4 = 5 = 6 = 7$, putting $i = 4$ between 5 and 1 instead of above them. The diamonds (◊) shown for several cylinders in Figure 1 are not acceptably close to the circles. The calculated order of increasing log h ($=y_j$) is $7 < 3 < 10 < 4 < 9 < 1 < 2 < 6 < 8 < 5$ instead of the true order, $7 < 9 < 3 < 10 < 1 < 2 < 8 < 4 < 5 < 6$. With $a_7 = -a_1$ as the sixth subsidiary condition, the calculated order of increasing log h is still different, namely, $7 < 4 < 10 < 2 < 3 < 9 < 6 < 1 < 8 < 5$. Such results based on accurate but incomplete data could be dangerously misleading in real problems, and therefore worse than useless.

Although SPSS does not provide convenient facilities for substituting the mean of data for an i in place of missing data for that i, use of means is claimed by many authors to be preferable to pairwise deletion. Column 7 shows the result (for $a_3 = a_1/2$). The correlation coefficient between observed and calculated data for the 50 is now 0.813, not much different with this example than pairwise deletion. The a_i parameters differ by more than 50% for $i = 2$, 4, and 6, although these should all be identical. A partial rank order for log h is $1 < 3 < 10 < 4 < 2 < 6 < 8 < 5$ vs. the true order, $3 < 10 < 1 < 2 < 8 < 4 < 5 < 6$. The percent of the variance explained by increasing numbers of modes is 68.5 (1), 82.9 (2), 94.2 (3), 97.8 (4), 99.4 (5), 99.9 (6), and 100.0 (7), which would mislead us into believing that at least three modes are involved and should be considered, *if we did not know that there are only two.*

In a relatively expensive but still futile attempt to find a practical way to obtain correct parameters from these 50 accurate data by factor analysis, we resorted to iteration of the combination of principal components extraction plus the complete transformation described above. Missing data for any i were replaced by the mean of data for that i before extraction in the first cycle, but by data predicted from the latest parameters (after transformation) in each subsequent cycle. This iterative procedure has serious disadvantages compared with DOVE. First, it expands each multiple regression to involve the latest estimates of all the missing data as well as the measured data, increasing the number of calculations involved in the time-consuming summations by 40% in the present example, and fourfold in our solvent effect problem where 75% of the data are missing. Second, although most of the parameters appear to be converging, it still gives a serious number of bad parameter values after five iterations (after ten job steps and computer charges ten times those required to obtain correct data and parameters to more than six decimal places by DOVE). At this point, the % variance explained has leveled off, but the correlation coefficient has begun to drop, as shown in columns 7–10 of Table I. Lest anyone believe that these parameters are now nevertheless good enough because 98.7% of the variance is explained, we should note that the correlation coefficient is only 0.917 and that the fit to the observed data is becoming worse, that cylinder 1 is still deduced to be shorter than cylinder 10 ($\log h_1 < \log h_{10}$), whereas it is actually taller, the sensitivities to $\log h$ appear to vary more than 50% for properties such as 2, 4, 6, and 7, for which they are truly identical, and those for 4, 5, and 6 are getting worse. Doubtless there would be other inversions in rank orders if the distribution of the missing data were more biased (more associated with particular i's and j's), as they often are in real problems, rather than being random and relatively uniform as in this example. The squares (□) in Figure 1 show that b_2 parameters for $i = 4$ and 6 are bad and currently becoming worse. It would be interesting to follow this progression to its limit, where all values might possibly end up correct, but they are changing too slowly to make this practicable.

From these and other experiences with iterative factor analyses, we have found that it is critically important not to rely on convergence or constancy of any missing datum, parameter, % variance, or correlation coefficient as a criterion of correctness, but to test any proposed program by a problem like the present example that has its answers known in advance.

We have not found any published literature documenting a claim that correct results have ever been obtained by an application of principal components or any other correlation coefficient-based factor analysis to data sets with missing data. We conclude that although DOVE is able to solve such problems by simply omitting missing data, standard factor analyses can neither omit them nor obtain satisfactory estimates of the missing data or parameters by any efficient successive approximation process.

CONCLUSION

There are two hazards in factor analysis. First, correct parameters cannot be obtained by principal components or any other kind of factor analysis based on correlation coefficients when data are missing. Usually the best way to obtain correct parameters is to measure the missing data to obtain a full data set, or narrow the scope of the study to a full subset having no missing data and then use principal components analysis followed by a valid transformation. If neither of these is practicable, we recommend use of the iterative DOVE procedure illustrated in the previous paper, which works correctly in spite of missing data.

Second, even with no missing data, the resulting parameters (factor loadings and factor scores) have no simple meaning unless the required number of valid subsidiary conditions is incorporated into a parameter transformation. The varimax or other standard rotations, which have always been used instead in the past, incorporate generally untrue conditions and are likely to yield parameters suggesting absurd or highly misleading conclusions.

Reproduction of the data is not a sufficient test that a procedure is correct, because an infinite number of sets of answers (parameters) are consistent with the data. It is indispensible to test any new or untested procedure on a problem having the same number of modes and with answers known in advance, because errors in assumptions or logic are almost certain to escape detection if only real problems with unknown answers are analyzed.

ACKNOWLEDGMENT

We thank Professor Albert J. Petkau, Department of Mathematics, University of British Columbia, for discussions and many helpful suggestions. This work was supported in part by the Office of Naval Research and the Petroleum Research Fund of the American Chemical Society. Mr. Bryndza acknowledges a Hertz Foundation Scholarship and computer charge credits by the MIT Student Information Processing Board.

REFERENCES AND NOTES

(1) This article focuses on 1974–1977 studies using standard or conventional methods of factor analysis, especially the relatively simple and popular method usually called principal components. Cf. ref 2, which illustrates a practical alternative approach capable of yielding correct parameters even when data are missing.

(2) P. F. Strong, C. G. Swain, and M. S. Swain, preceding paper in this issue.

(3) Beginning with K. Pearson, *Phil. Mag.*, Ser. 6, **2**, 559 (1901), they include 530 before 1940 (D. Wolfle, *Psychometric Monogr.*, **3**, 44–66 (1940)), 786 for 1940–52(B. Fruchter, "Introduction to Factor Analysis", Van Nostrand, New York, 1954, pp 221–266)), and 640 even *excluding* psychology (in ref. 10, pp 522–600) for 1952–66.

(4) R. N. Kavanagh, T. M. Darcey, and D. H. Fender, *Electroencephalogr. Clin. Neurophysiol.*, **40**, 633 (1976); C. S. Newmark, T. R. Faschingbauer, A. J. Finch, and P. C. Kendall, *J. Clin. Psychol.*, **31**, 449 (1975); D. J. Tosi and D. M. Eshbaugh, *ibid.*, **32**, 322 (1976).

(5) H. A. Guthrie and G. M. Guthrie, *Am. J. Clin. Nutr.*, **29**, 1238 (1976); A. Cammarata and G. K. Menon, *J. Med. Chem.*, **19**, 739 (1976); M. L. Weiner and P. H. Weiner, *ibid*, **16**, 655 (1973).

(6) W. J. Moore, *J. Anat.*, **123**, 111 (1977); S. L. Horowitz, B. Graf-Pinthus, M. Bettex, H. Vinkka, and L. J. Gerstman *Arch. Oral Biol.*, **21**, 465 (1976); A. M. Henderson and D. L. Greene, *J. Dent. Res.*, **54**, 344 (1975); A. V. Lombardi, *Am. J. Phys. Anthropol.*, **42**, 99 (1975); D. G. Barker, *ibid.*, **44**, 27 (1976); E. H. Bryant and W. R. Atchley, Ed., "Benchmark Papers in Systematic and Evolutionary Biology", No. 2, Wiley, New York, N.Y., 1975, pp 1–345; G. Darland, *Appl. Microbiol.*, **30**, 282 (1975); C. S. Moore, *Ann. Bot.* (*London*), **39**, 113 (1975); M. V. Angel and M. J. R. Fasham, *J. Mar. Biol. Assoc. U.K.*, **55**, 709 (1975); H. J. B. Birks and M. Saarnisto, *Boreas* (*Oslo*), **4**, 77, (1975); J. M. Perkins, *Heredity*, **32**, 189 (1974); R. A. Orth, L. O'Brien, and R. Jardine, *Aust. J. Agric. Res.*, **27**, 575 (1976); L. S. Wu, R. E. Bargmann, and J. J. Powers, *J. Food Sci.*, **42**, 944 (1977).

(7) P. H. Weiner, "Solve Problems via Factor Analysis", *Chemtech.*, **7**, 321–328 (1977), 77 references; M. Bohle, W. Kollecker, and D. Martin, "Use of Factor Analysis in Organic Chemistry", *Z. Chem.*, **17**, 161–168 (1977), 51 references; D. G. Howery, *Am. Lab.*, **8**(2), 14 (1976), 32 references; R. W. Rozett and E. M. Petersen, *ibid.*, **9** (2), 107–116 (1977), 25 references; *Anal. Chem.*, **48**, 817–825 (1976), 29 references; J. R. McGill and B. R. Kowalski, *ibid.*, **49**, 596–602 (1977); G. L. Ritter, S. R. Lowry, T. L. Isenhour, and C. L. Wilkins, *ibid.*, **48**, 591 (1976); J. B. Justice, Jr., and T. L. Isenhour, *ibid.*, **47**, 2286 (1975); R. N. Carey, S. Wold, and J. O. Westgard, *ibid.*, **47**, 1824 (1975); J. E. Davis, A. Shepard, N. Stanford, and L. B. Rogers, *ibid.*, **46**, 821 (1974); J. T. Bulmer and H. F. Shurvell, *Can. J. Chem.*, **53**, 1251 (1975); E. R. Malinowski, *Anal. Chem.*, **49**, 606, 612 (1977); R. N. Cochran and F. H. Horne, *ibid.*, **49**, 846 (1977).

(8) P. K. Hopke, E. G. Gladney, G. E. Gordon, W. H. Zoller, and A. G. Jones, *Atmos. Environ.*, **10**, 1015–1025 (1976); L. G. Closs and I. Nichol, *Can. J. Earth Sci.*, **12**, 1316 (1975); H. Hoetzl, *Dtsch. Geol. Ges. Z.*, **126**, 121 (1975); E. L. Zodrow, *J. Int. Assoc. Math. Geol.*, **8**, 395 (1976); A. I. Benimoff and N. E. Pingitore, Jr., *Neues Jahrb. Mineral., Monatsh.*, **84** (1977); J. S. Duval, *Geophysics*, **42**, 549 (1977); A. T. Miesch, *Geol. Surv. Prof. Pap.* (*U.S.*), 574-G (1976); N. E. Pingitore, Jr., and J. D. Shotwell, *Neues Jahrb. Geol. Palaeontol., Monatsh.*, 177 (1977).

(9) J. E. Vincent, "Factor Analysis in International Relations", University of Florida Press, Gainesville, Fla., 1971; D. R. Hall and R. J. Rummel, "The Dimensionality of Nations Project", Research Report No. 16, University of Hawaii, 1969; C. Wall and R. J. Rummel, *ibid.*, Report No. 20 (1969); B. M. Russett, Ed., "Peace, War, and Numbers", Sage, Beverly Hills, Calif., 1972; P. Warwick, *Soc. Sci. Res.*, **14**, 241 (1975).

(10) R. J. Rummel, "Applied Factor Analysis", Northwestern University Press, Evanston, Ill., 1970.

(11) D. Child, "Essentials of Factor Analysis", Holt, Rinehart, and Winston, London 1970; R. L. Gorsuch, "Factor Analysis", Saunders, Philadelphia, Pa., 1974; H. H. Harman, "Modern Factor Analysis", 3rd ed, University of Chicago Press, Chicago, Ill., 1976; A. L. Comrey, "A First Course in Factor Analysis", Academic Press, New York, N.Y., 1973; S. A. Mulaik, "Foundations of Factor Analysis", McGraw-Hill, New York,

N.Y., 1972; W. H. Guertin and J. P. Bailey, Jr., "Introduction to Modern Factor Analysis", Edwards, Ann Arbor, Mich., 1970; D. N. Lawley and A. E. Maxwell, "Factor Analysis as a Statistical Method", American Elsevier, New York, 1971; W. W. Cooley and P. R. Lohnes, "Multivariate Data Analysis", Wiley, New York, 1971, pp 96–167; D. F. Morrison, "Multivariate Statistical Methods", 2nd ed, McGraw-Hill, New York, 1976, pp 266–343; A. S. C. Ehrenberg, "Data Reduction", Wiley, New York, 1975, pp 263–282; K. Enslein, A. Ralston, and H. S. Wilf, Ed., "Statistical Methods for Digital Computers", Vol. 3, Wiley, New York, 1977, pp 9–10, 123–202; E. Weber, "Einführung in die Faktorenanalyse", VEB Gustav Fischer Verlag, Jena, 1974.

(12) N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent, "Statistical Package for the Social Sciences", 2nd ed, McGraw-Hill, New York, 1975, pp 468–514.

(13) W. J. Dixon, "BMD Biomedical Computer Programs", University of California Press, 1973; W. J. Dixon, "BMPD Biomedical Computer Programs", University of California Press, 1975.

(14) D. G. Armor and A. S. Couch, "Data-Text Primer", Macmillan, New York, 1972, pp 82–91.

(15) R. Buhler, "P-STAT", Princeton University, 1975.

(16) "System/360 Scientific Subroutine Package Programmer's Manual, Version III", International Business Machines Corp., Manual GH20-0205-4, 1968.

(17) "IMSL Library 1" (IBM Series), 5th ed, 1975; and "IMSL Library 2" (Honeywell Series), 4th ed, 1974, International Mathematical and Statistical Libraries, Houston, Texas.

(18) B. H. Hall, "Time Series Processor, Version 2," Harvard Institute of Economic Research, Harvard University, Cambridge, Mass., 1976.

(19) L. L. Thurstone, "Multiple-Factor Analysis", University of Chicago Press, Chicago, Ill., 1947, pp 117–124.

(20) The eight synthetic problems or "artificial experiments" listed in ref 10, pp 528–9.

(21) Four SPSS extraction procedures (PA1, PA2, Rao, and alpha) with six rotations (varimax, equimax, quartimax, and oblique with $\delta = -1$, 0, +1) were carried out. Image extraction failed (singular correlation matrix) even with 0.1–1% data errors (absolute magnitude changed by 1 in third significant figures, up for odd $i + j$, down for even $i + j$). None of the combinations gave realistic sets of factors. Quartimax also made b_2 positive and b_7 negative. Highly oblique ($\delta = +1$) structure factors were only slightly negative for b_2 but varied more than threefold from b_4 to b_7. Most other sets had numerous incorrect signs.

(22) One-mode problems, such as the chemical one of best fitting the Hammett equation with a single type of substituent constant (S. Wold and M. Sjöström, *Chem. Scr.*, **2**, 49 (1972); S. Wold, *ibid.*, **5**, 97 (1974); M. Sjöström and S. Wold, *ibid.*, **6**, 114 (1974); **9**, 200 (1976)), do not lead to such errors and confusion, because the number of critical conditions is zero for them; hence there is no possibility of choosing the critical conditions incorrectly.

Graph-Based Fragment Searches in Polycyclic Structures

MILAN RANDIĆ*[1] and CHARLES L. WILKINS*

Department of Chemistry, University of Nebraska—Lincoln, Lincoln, Nebraska 68588

Received May 31, 1978

A procedure for substructure search in polycyclic systems based upon the concept of atom codes and their comparison is outlined. The atom codes used represent the number of paths of different length originated at each atom. This scheme represents an extension to polycyclic structures of a recently described algorithm for fragment searches in acyclic structures. The basic idea of using the number of neighbors and number of paths for characterization of an atomic environment is not unfamiliar. However, it appears that the accompanying formalism has not been developed previously. The path concept has found much application in various problems relating graphs and structures, but previously the primary emphasis has been the search for shortest paths. Here we show it is useful to list all paths (to any desired depth) for all atoms in a structure and then to use such a list for substructure searches. Essentially, comparison of the lists of paths for two graphs permits establishing compatibilities among the vertices and results in assignment of multiple labels to the graph under investigation, where the labels are those of the fragment. Subsequent atom-by-atom matching is straightforward and efficient, since the use of multiple labeling prevents a large number of otherwise unproductive searches from even being considered. By successive registration of a fragment, vertices whose labels have been exhausted are deleted and the process repeated on a smaller graph with fewer multiple labels—which makes possible rapid convergence in the search. This search procedure is illustrated with an example.

INTRODUCTION

The problem of fragment searching or, as it is known among mathematicians, the problem of subgraph isomorphism, continues to attract considerable attention in diverse scientific disciplines. It appears under different guises in its special forms (e.g., the search for the shortest path or a search for a Hamiltonian circuit). In chemistry, the problem is of interest not only to those involved in chemical documentation and the use of computers in searching for optimal synthetic routes but also has possible implications for pattern recognition and other manipulations utilizing large files of data, as well as for drug design and/or structure–activity correlations. The continuing development of experimental techniques and the rapid accumulation of more reliable and more accurate experimental data mandates improved algorithms for computer manipulation of the results.

Comparison of structural representations and their correlation with associated data requires efficient fragment search algorithms. Here we are primarily interested in substructure searches, where the term implies a collection of connected atoms, rather than searching for a substructure identified by a conventional (or trivial) name such as is used in substructure searches of an index, for example. The input information in our case is the connection table for atoms forming the fragment, or pictorially, the graph corresponding to the fragment. The problem of subgraph isomorphism has been considered in the literature, though not so thoroughly as the related problem of graph isomorphism.[2] Conceptually, the simplest and probably the oldest scheme is atom-by-atom matching which is equally suitable and, in application, practically identical whether one searches for graph isomorphism or subgraph isomorphism.[3] In an atom-by-atom search, one proceeds from the selected atom to its neighbors sequentially selecting single bonds and continues the process as long as matches are found. As soon as a nonmatch occurs, one back-tracks to the previous branching point and selects an alternative bond. The search is continued until the desired fragment is found or until all alternative routes have been exhausted. Besides requiring back-tracking, which is wasteful, the scheme also requires extensive bookkeeping, and even for

0095-2338/79/1619-0023$01.00/0 © 1979 American Chemical Society