Methods in Geomathematics 1

GEOLOGICAL FACTOR ANALYSIS

by

K.G. JÖRESKOG J.E. KLOVAN R.A. REYMENT

Methods in Geomathematics 1

GEOLOGICAL FACTOR ANALYSIS

by

K.G. JÖRESKOG

Department of Statistics, University of Uppsala, Uppsala, Sweden

J.E. KLOVAN

Department of Geology, University of Calgary, Calgary, Alta., Canada

R.A. REYMENT

Department of Historical Geology and Paleontology, University of Uppsala, Uppsala, Sweden

ELSEVIER SCIENTIFIC PUBLISHING COMPANY
Amsterdam — Oxford — New York 1976

ELSEVIER SCIENTIFIC PUBLISHING COMPANY
335 Jan van Galenstraat
P.O. Box 211, Amsterdam, The Netherlands

AMERICAN ELSEVIER PUBLISHING COMPANY, INC.
52 Vanderbilt Avenue
New York, New York 10017

ISBN 0-444-41367-7

With 53 illustrations and 45 tables

Printed in The Netherlands

PREFACE

Our decision to write this book stems from the fact that the kind of analysis involved in what we group under the heading of "Geological Factor Analysis" has become one of the most frequently used sets of multivariate statistical techniques in geology and the general concepts of which many geologists have at least a vague understanding. In putting all of these techniques into the same bag, we recognize the fact that the term "factor analysis" has come to be applied by geologists to a particular kind of analytical procedures of which only a few belong to the classical factor model such as it is conceived by psychometricians.

It is the aim of our text to introduce students of geology to the powerful technique of factor analysis and to provide them with the background necessary in order to be able to undertake analyses on their own. For this reason, we have tended perhaps to be over-explicit when dealing with the introductory requirements for understanding the calculations and to have paid less attention to theoretical details. Clearly, we have definitely not written a text for statisticians.

The analysis of homogeneous multivariate populations in the earth sciences has grown into a primary research branch of almost unlimited potential; this development, largely made possible by the rise of the electronic computer, has greatly altered methodology in the petroleum industry, mining geology, geochemistry, stratigraphy, palaeontology, chemical geology, environmental geology, sedimentology and petrology.

Chapter 1 introduces the concept of multivariate data analysis by factoring methods. In Chapter 2, we present the basic concepts of multivariate algebra (linear algebra, matrix algebra) and the most commonly occurring matrix arithmetic operations of factor analysis. We also consider the rotation of coordinate systems and the role of eigenvalues and eigenvectors.

In Chapters 3 and 4, we take up theoretical concepts of factor analysis and the statistical interpretation of models, in which the presentation is made in terms of fixed mode and random mode: we give an account of the methods of principal components, and "true" factor analysis. In Chapter 5, Q-mode factor analysis, principal coordinates and correspondence analysis are presented. Chapter 6 is concerned with every-day practical problems you are liable to run into when you plan and carry out factor analyses. Here, such diverse topics as selection of the most suitable method, choosing the number of meaningful factors, and data transformations, are discussed.

Chapter 7 takes you through examples of each of the major techniques. We end up this section with a set of reviews of randomly chosen applications of factor analysis from the literature.

It is hoped that this book, the result of the collaboration between a professional statistician and two geologists with experience in various fields of application of the methods presented here, will prove useful to students and research workers alike.

It is expected that most users of our text will have little knowledge of the general field of statistics and we have, therefore, chosen to develop our subject at an elementary level. We wish, nevertheless, to recommend strongly to those who lack a background in statistics to do some introductory reading in the subject.

The opportunity of testing our approach to Factor Analysis was given to us in the Spring of 1974, when we gave a post-graduate course on the subject at Uppsala University. This cooperation at the teaching level provided a valuable sequel to the year of preparation behind the book and helped to iron out difficulties concerning presentation of the text. Thanks to a grant from the Natural Research Council of Canada, Klovan was able to spend the greater part of the Academic year of 1973—1974 at Uppsala.

Several colleagues have aided in furthering the development of the book. In particular, we wish to mention Mr. Hans-Åke Ramdén, Uppsala Datacentral, Mr. John Gower, Rothamsted Experimental Station, U.K., Dr. M. David, Ecole Polytechnique, Montréal, Canada, Mr. M. Hill, Natural Environment Research Council, Bangor, U.K., Dr. A.T. Miesch, U.S. Geological Survey, Denver, U.S.A., and Mr. Colin Banfield, Rothamsted, U.K.

K.G. JÖRESKOG, J.E. KLOVAN, R.A. REYMENT

GLOSSARY OF THE MOST COMMONLY USED SYMBOLS

N = the number of objects (specimens, observations) in a sample; it denotes the size of the sample
p = the number of variables (characters, attributes)
k = the number of factors
$\mathbf{X}$ = the data matrix (the order of which is $N \times p$)
$\mathbf{Z}$ = the standardized data matrix
$\mathbf{W}$ = the row-normalized data matrix of Imbrie Q-mode factor analysis
$\mathbf{R}$ = the sample correlation matrix
$\mathbf{S}$ = the sample covariance matrix
Σ = the population covariance matrix
$\mathbf{H}$ = the association matrix of Q-mode methods
Λ = the population and sample diagonal matrix of eigenvalues; the elements of this matrix are λ_i
$\mathbf{U}$ = the sample matrix of eigenvectors
$\mathbf{F}$ = the matrix of factor scores
$\mathbf{A}$ = the matrix of factor loadings
$\mathbf{E}$ = the matrix of residuals or error terms
i and j are used to denote the indices for rows and columns of a $p \times p$ matrix (for example, the correlation matrix)
m and n are used to denote row and column indices for $N \times N$ matrices; the data matrix $\mathbf{X}_{(N \times p)}$ has then a general element x_{mj}, a correlation matrix $\mathbf{R}$, a general element r_{ij} and an association matrix $\mathbf{H}$, a general element h_{mn}

CONTENTS

Chapter 1

INTRODUCTION

1.1 STRUCTURE IN MULTIVARIATE DATA

Commonly, almost all geologists make a great number of measurements in their daily activities, for example, on the orientation of strata, geochemical determinations, mineral compositions, rock analyses, measurements on fossil specimens, properties of sediments, and many other kinds. You need only reflect on the routine work of a geological survey department in order that the truth of this statement may become apparent.

The data of geology are very often multivariate. For example, in a rock analysis, determinations of several chemical elements are made on each rock specimen of a collection. You will all be familiar with the tables of chemical analyses that issue from studies in igneous petrology. Petrologists have devised many kinds of diagrams in their endeavour to identify significant groupings in these data lists. The familiar triangular diagrams of petrology permit the relationships between three variables at a time to be displayed. Attempts at illustrating more highly multivariate relationships have led to the use of ratios of elements and plots on polygonal diagrams.

Obviously, one can only go so far with the graphical analysis of a data table. The logical next step is to use some type of quantitative method for summarizing and analyzing the information hidden in a highly multivariate table. It is natural to enquire how the variables measured for a homogeneous sample are connected to each other and whether they occur in different combinations, deriving from various relationships in the population. One may, on the other hand, be interested in seeing how the specimens or objects of the sample itself are inter-related, with the thought in mind of looking for natural groupings. In both cases, we should be looking for structure in the data.

Geologists came into touch with the concept of factor analysis and the study of multivariate data structure through the contacts between palaeontologists and biologists. The biologists, in their turn, learnt the techniques from psychometricians. Thus, the French zoologist Teissier studied multivariate relationships in the carapace of a species of crabs (Teissier, 1938), using a centroid first-factor solution of a correlation matrix. He interpreted this "general factor" as one indicating differential growth.

Let us now look briefly at a few typical problems that may be given meaningful solutions by an appropriately chosen variety of factor analysis.

A geochemist has analyzed several trace elements in samples of sediment from a certain area and he wishes to study the relationships between these elements in the hope of being able to draw conclusions on the origin of the sediment.

A mining geologist is interested in prospecting an area for ores and wants to use accumulated information on the chemistry and structural geology of known deposits in the region to help predict the possibilities of finding new ore bodies.

A palaeontologist wishes to analyze growth and shape variation in the shell of a species of brachiopods on which he has measured a large number of characters.

A petroleum company wants to reduce the voluminous accumulations of data deriving from palaeoecological and sedimentological studies of subsurface samples to a form that can be used for exploring for oil-bearing environments.

In an oceanological study, it is desired to produce graphs which will show the relationships between bottom samples and measurements made upon them on a single diagram as a means of relating organisms to their preferences for a particular kind of sediment.

1.2 AN EXAMPLE OF FACTOR ANALYSIS

At this point, we think it would be helpful to you if we gave you an inkling of what is obtained in a factor analysis. We have chosen an albeit artificial mining example by Klovan (1968) as it not only introduces the geological element at an early stage but also because it does provide a good practically oriented introduction to the subject.

Imagine the following situation. We wish to carry out exploration for lead and zinc in an area containing a high-grade lead-zinc ore. The area has been well explored geologically and the bedrock is made up of an altered carbonate-shale sequence. The map area and the sampling grid are displayed in Fig. 1.1.

The three controls, palaeotemperature (T), strength of deformation of the bedrock (D), and the permeability of the rock (P) are considered to determine

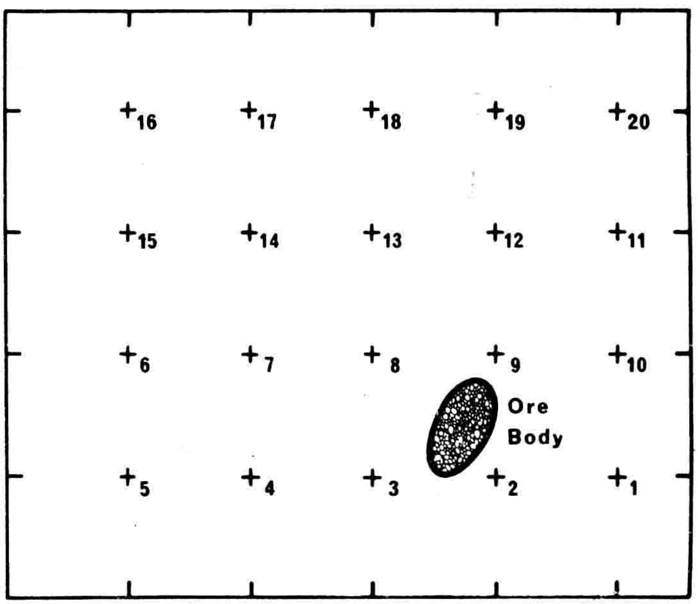
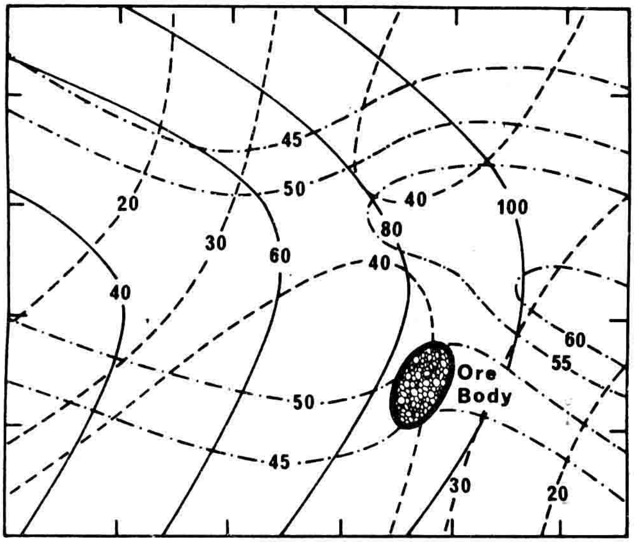


Fig. 1.1. The sampling grid for the prospecting example.



——— DISTRIBUTION OF PALEOTEMPERATURES

—·—·— DISTRIBUTION OF PERMEABILITY

- - - - - DISTRIBUTION OF DEFORMATION

Fig. 1.2. Distribution of controls imposed at the locations of the samples.

the occurrence of lead and zinc, for the purposes of our example. It is assumed that these controls are determinable from observations on ten chemical, mineralogical and rock-deformational variables. The distribution of these causes will in reality never be known but, for this example, we shall imagine that they are distributed as shown in Fig. 1.2. You will note that the lode lies at the intersection of these causes at certain specified levels. These are, for palaeotemperature, 80—90, for deformation, 35—45, and for permeability of the country rock, 45—50. Accepting that a geological survey of the area would have given as clear an indication as our manufactured example, it would not be unreasonable to expect that target areas for intensive prospecting would occur in localities where the intersection situation is repeated.

The three controls can, of course, not be estimated directly. They can, however, be measured indirectly from geological properties that are a reflection of them. The arrays shown in Table 1.I list the artificial data, as well as the information used in constructing this set of observations. The left array of numbers gives the "amount" of each of the three controls at each of the localities; the upper array states precisely the degree to which each of the geological variables is related to the causes. Multiplication and summation of every row of the left array by every column of the top array yields the large array (corresponding to raw data) at the bottom. Naturally, in a real study, you would not know the left-hand and top arrays of Table 1.I. All you would have at your disposal would be the large array, or data matrix, the result of a detailed geological survey and a laboratory study of the samples collected.

The question to be answered now is, how can we determine the existence of

TABLE 1.I.

THE RAW DATA MATRIX FOR THE LEAD—ZINC PROSPECTING PROBLEM

				Geological properties									
				Mg in calcite	Fe in sphalerite	Na in muscovite	Sulphide	Crystal size of carbonates	Spacing of cleavage	Elongation of ooliths	Tightness of folds	Vein material per metre^2	Fractures/metre^2
		Causes	*T*	0.95	0.75	0.75	0.33	—0.20	0.05	0.20	0.10	0.00	0.05
			D	0.00	0.10	0.20	0.33	0.60	0.95	0.70	0.85	0.10	0.25
			P	0.05	0.15	0.05	0.34	0.60	0.00	0.10	0.05	0.90	0.70
Locality	*T*	*D*	*P*	Data matrix									
1	121	21	46	1175	999	975	625	158	262	437	324	431	433
2	96	35	42	936	820	813	575	267	379	478	413	411	428
3	78	54	49	765	711	716	599	457	548	579	558	491	513
4	63	51	49	624	598	600	542	471	515	531	520	490	500
5	42	44	44	417	422	422	432	444	441	437	439	437	437
6	39	26	54	401	403	375	401	405	270	317	290	515	465
7	52	36	52	520	504	488	469	427	370	410	386	507	482
8	67	46	54	661	626	618	553	462	466	506	480	529	523
9	90	37	51	877	787	773	594	354	401	493	434	500	498
10	108	27	61	1060	932	898	656	315	312	468	370	580	552
11	112	33	59	1010	960	935	681	334	375	518	427	567	555
12	91	38	59	896	811	790	629	403	411	511	448	570	555
13	76	39	54	748	688	672	560	401	399	472	426	525	512
14	63	30	51	617	573	553	477	360	315	385	342	487	462
15	43	19	55	436	424	389	393	361	207	277	236	514	455
16	68	16	42	664	587	560	419	212	182	287	221	397	369
17	77	27	41	750	665	651	484	259	299	387	331	399	396
18	93	37	43	903	797	791	573	291	396	486	427	421	437
19	102	47	48	998	888	887	657	366	499	583	527	480	506
20	120	36	46	1162	999	994	671	252	404	539	450	449	471

structure in such a large array of numbers? The technique of factor analysis turns out to be a useful way of providing plausible answers.

Simply put, factor analysis creates a minimum number of new variables which are linear combinations of the original ones such that the new variables contain the same amount of information.

The starting point is provided by the correlations between the variables measured, ten in all. The matrix of correlation coefficients is listed in Table 1.II. It was subjected to principal components factor analysis for which three significant factors were obtained. Thus, we began with ten characters but can now "explain" the total variability of the sample in terms of three new variables or factors.

The principal-factor matrix is listed in Table 1.III; it shows the "composition" of the factors in relation to the original variables. As these factors are usually not readily interpretable, it is accustomed practice to rotate the reference axes

TABLE 1.II

INTER-CORRELATIONS AMONG THE TEN GEOLOGICAL PROPERTIES

	Mg	Fe	Na	S	crystal	cleavage	ooliths	folds	veins	fractures
1	1.000									
2	0.998	1.000								
3	0.994	0.998	1.000							
4	0.908	0.933	0.942	1.000						
5	−0.576	−0.523	−0.497	−0.180	1.000					
6	0.130	0.183	0.235	0.477	0.616	1.000				
7	0.581	0.625	0.664	0.834	0.258	0.880	1.000			
8	0.282	0.334	0.383	0.610	0.519	0.987	0.944	1.000		
9	0.012	0.057	0.035	0.286	0.539	0.181	0.216	0.208	1.000	
10	0.258	0.313	0.312	0.590	0.550	0.524	0.604	0.573	0.909	1.000

by some appropriate method in order to bring out the important contributing loadings and to diminish the loadings on non-significantly contributing variables. The visual result of the rotation will then be that some of the loadings will have been augmented while others will have become greatly lower. In our example, we used the varimax rotation technique. The varimax factor matrix displayed in Table 1.III demonstrates what we have just described and you will see this if you compare entries in the two upper listings of the table, entry by entry. The rotated factor matrix contains ten rows and three columns, each latter representing a factor. Reading down a column, the individual numbers tell us the contribution of a particular variable to the composition of the factor; in fact, each column can be thought of as a factor equation in which each loading is the coefficient of the corresponding original variable.

A third chart of numbers emerges from the factor analysis, the varimax factor score matrix, shown in Table 1.III. This gives the amounts of the new variables at each of the sample localities. With this matrix, we are able to map the distributions of these new factor variables on the sample grid.

It requires sound geological reasoning in order to interpret the results of a factor analysis. From Table 1.III, you will see that the first factor is mainly concerned with the variables: Mg in calcite, Fe in sphalerite, and Na in muscovite, a combination indicating temperature dependence. The second factor is heavily loaded with the variables: spacing of cleavage, elongation of ooliths, and tightness of folds, a combination speaking for rock deformation. The third factor is dominated by the variables: vein material per m^2, and fractures per m^2, interpretable as being a measure of permeability of the country rock.

The distribution of the three sets of factor scores is shown in Fig. 1.3. The patterns of Fig. 1.2 are almost exactly duplicated. By comparing the nature of the intersections around the known ore body, and searching the diagram for a similar pattern, you will see that at least one other area on the map has the

TABLE 1.III

RESULTS OF THE FACTOR ANALYSIS

Variable	Communality	Factors		
		1	2	3
Principle factors of correlation matrix				
1	1.0000	0.8029	−0.5894	0.0886
2	1.0000	0.8385	−0.5367	0.0940
3	1.0000	0.8579	−0.5122	0.0407
4	1.0000	0.9760	−0.1961	0.0943
5	1.0000	0.0176	0.9998	−0.0098
6	1.0000	0.6538	0.5999	−0.4611
7	1.0000	0.9297	0.2393	−0.2799
8	1.0000	0.7647	0.5018	−0.4042
9	1.0000	0.3268	0.5407	0.7751
10	1.0000	0.6641	0.5437	0.5132
Variance		54.614	31.928	13.459
Cumulative variance		54.614	86.542	100.000
Varimax factor matrix				
1	1.0000	0.9971	0.0765	−0.0060
2	1.0000	0.9916	0.1241	0.0362
3	1.0000	0.9835	0.1804	0.0117
4	1.0000	0.8813	0.3985	0.2540
5	1.0000	−0.6197	0.5880	0.5198
6	1.0000	0.0558	0.9897	0.1317
7	1.0000	0.5191	0.8380	0.1680
8	1.0000	0.2102	0.9648	0.1580
9	1.0000	0.0146	0.0488	0.9987
10	1.0000	0.2338	0.3979	0.8872
Variance		44.771	33.318	21.912
Cumulated variance		44.771	78.089	100.000

Varimax factor score matrix

Locality	Factors		
	1	2	3
1	1.7291	−1.1345	−0.9480
2	0.6130	0.2141	−1.3682
3	−0.2291	1.8610	0.0226
4	−0.7962	1.5403	0.0196
5	−1.6301	0.9332	−0.8890
6	−1.5345	−1.0766	0.6182
7	−1.1157	−0.0212	0.4175
8	−0.5908	0.9143	0.7663
9	0.3711	0.2439	0.2588
10	1.2434	−0.9398	1.7595
11	1.3197	−0.2480	1.4876
12	0.4639	0.1764	1.5324
13	−0.1684	0.1884	0.7211
14	−0.6634	−0.5747	0.0782
15	−1.3364	−1.7591	0.6394
16	−0.3829	−1.7847	−1.5171
17	−0.1144	−0.5587	−1.5440
18	0.4618	0.3838	−1.1944
19	0.7982	1.3086	−0.1605
20	1.5620	0.3336	−0.6997

same special conditions. The marked square is thus the first-order target for further exploration. This is an artificial example, contrived to give a good result. Under actual exploration conditions, you would not expect things to fall out so nicely and your geological knowledge would be put to a greater test.