

Computational Methods of Multivariate Analysis in Physical Geography

P.M. Mather

57.12
M427

*Computational Methods
of
Multivariate Analysis
in
Physical Geography*

P. M. Mather

*Lecturer in Geography,
University of Nottingham*

JOHN WILEY & SONS

London • New York • Sydney • Toronto

Copyright © 1976, by John Wiley & Sons.

All rights reserved.

No part of this book may be reproduced by any means, nor transmitted, nor translated into a machine language without the written permission of the publisher.

Library of Congress Cataloging in Publication Data:

Mather, Paul M.

Computational methods of multivariate analysis in physical geography.

1. Physical geography — Mathematics. 2. Multivariate analysis. I. Title.

GB21.M37 910'.02'051953 75-23376

ISBN 0 471 57626 3

Typeset in IBM Baskerville by Preface Ltd., Salisbury, Wilts., and printed by The Pitman Press Ltd., Bath

Preface

This book is the outcome of my interest in the use of multivariate analysis in physical geography. One of my earliest impressions of life as a research student in the mid-1960's was the lack of textbooks on advanced statistical and data-processing techniques written by geographers for geographers. Consequently, I had to rely on books such as Rao (1952) and Cattell (1966a) which, although excellent in themselves, required an appreciation of the terminology and research problems of psychometricians and biometricians before the methods themselves could be understood. The appearance of L. J. King's book in 1969 went some way towards relieving this situation, but the level of technical detail was not high. It is still true to say that there is no geographical textbook which adequately sets out the technical basis of the more important multivariate techniques. By 'technical basis' I mean details of the specifications of the statistical models, with information on the assumptions of these models and discussion of the consequences resulting from their violation. I would also include description of alternative methods of estimation when the basic assumptions do not hold, and consideration of the numerical and computational aspects of the particular method, including details of feasible algorithms. This book represents my effort to fill this gap in the literature. The level of mathematical and statistical sophistication of research workers in physical geography is now sufficiently high for a technical book such as this one to be acceptable without the need for large-scale examples and case studies. Such examples have their uses but I feel that, in a book of this nature, specific examples, requiring a reasonable level of knowledge of the research problem involved, would distract the reader from the fundamental points concerning the technical basis of each method. Consequently, the examples are deliberately kept small and simple. In this form, they also provide suitable test data sets for the computer programs and algorithms which are described in the text. It is emphasized that these examples are not meant to be representative

applications of the use of multivariate techniques in physical geography.

I anticipate that readers of this book will be postgraduate research workers in physical geography who are actively engaged in the analysis of problems that can be expressed quantitatively. Consequently, this book is not set out in the form of a course in multivariate analysis, but is written in the expectation that it will be used both as a reference book and as a source of computer programs. This approach justifies the concentration on technical and computational details rather than on the examples of the use of various techniques in a research context.

I would like to thank all those who have helped me in the preparation of this book. Professor J. P. Cole, Department of Geography, University of Nottingham, has been a constant source of ideas and enthusiasm. Mr. R. P. Bradshaw, of the Geography Department, University of Nottingham, has kindly commented on several sections of the manuscript. The staff of the numerical section of the Cripps Computing Centre, University of Nottingham, have provided invaluable advice and assistance. I am also grateful for the comments of an anonymous referee. The editorial staff of John Wiley and Sons, Ltd., have invariably been helpful, and extremely patient. Lastly, it would not have been possible to write this book without the patience and consideration of my wife.

Nottingham, 1975.

PAUL M. MATHER

Alphabetical List of Subroutines and Functions

Table numbers are indicated for each entry

ACCINV, A30	GJDEF1, A31	PLOTS, A8
ADDCOL, 3.8	GJDEF3, 5.19	PREP, 7.8
ADJUST, 5.7	GSORTH, A34	PREP1, 6.12
ALPHA, 5.3	HISTGM, A14	PRNTRI, A2
ARBOR, 6.12	HTDQL, A24	PROMAX, 5.15
BART, 5.19	IFIND, A19	RANDOM, A16
CANON, 5.7	IDLZD, 5.20	RANGE, 6.13
CANON, 7.8	IMAGE, 5.10	RANK, A13
CHEK, 6.13	IMPORT, 5.7	RATQR, A21
CHOLS, A29	INN, 6.14	REBAKA, A23
CLUMP, 6.13	INPT, 6.12, 6.16, 7.7	RECOV, 3.3, 3.8
CLUS, 6.12	KALK, A11	REDUC, A22
CLUSTR, 6.12	LINK, 6.12	REMOVE, 3.15
COMP, 3.3	MAP, 3.8	RESET, 3.15
COPHGN, 6.12	MATCH, 6.14	RESID, 2.1, 3.15
COPHCR, 6.12	MATPR, A1	REX, 2.1, 3.15, 4.2
COPY, A4	MAX, 3.15	SAMESC, A9
CROUT, A32	MDA, 7.8	SCALE, A10
DAPLOT, 3.8	MERGE, 6.13	SCORE, 4.2
DATIN, 2.1, 7.2	MIN, 6.12	SEARCH, 6.12
DERIVS, 6.16	MINIM, 6.12	SETUP, A15
DEVIAT, A18	MINRES, 5.7	SIGNIF, 2.1
DIST, 6.16	MONOT, 6.15	SIMCAL, 6.14
DISTPT, 6.15	MTMULT, A6	SMALL, 6.12
DNSAT1, 6.15	NONLIN, 6.16	SOLVE, A28
DPACC, A7	NORMLZ, 6.15	SORT, 6.15
DT2, A27	NSEIG, A33	SSCP, A3
EIGVEC, A25	ORTHCL, 3.3	SWAP, 5.15
EUCLID, 6.13	OUTPUT, 6.13	SWAP1, 7.7
FISHER, 3.15	OVCHK, A12	TARGET, 5.16
FPROB, 2.1	PCORD, 6.14	TRANS, 2.1
FSCORE, 5.18	PLACE, 7.8	TRBAK1, A26
FVAL, 5.7	PLOTIT, 3.3	TRIDI, A20

TRMULT, A5
TSA, 3.8
TSMAN, 3.8
UPSAT1, 6.15

VARIAN, 3.3
VARMAX, 5.12
VCALC, 3.15
XMOP, 3.3

YDIST, 6.16
ZED, 3.8
ZERO, A18

Contents

Introduction	1
0.0 Exploratory Data Analysis	1
0.1 Structure of this Book	3
0.2 Computer Programs	6

PART ONE ESSENTIALS OF MATRIX ALGEBRA

1 Mathematical Background	9
1.0 Introduction	9
1.1 Basic Definitions	10
1.2 Further Definitions	14
1.3 Vectors	16
1.4 The Inverse Matrix	23
1.5 Eigenvalues and Eigenvectors	28
1.6 Error in Matrix Computations	32
Appendix Summary of Matrix Theory	35

PART TWO ANALYSIS OF DEPENDENCE

2 Regression Analysis: the Linear Model	39
2.0 Introduction	39
2.1 The Linear Regression Model	40
2.2 Estimation in the Linear Regression Model	45
2.3 Computational Aspects of Multiple Regression	52
2.4 Problems in Regression	73
(i) Multicollinearity and Ridge Regression	73
(ii) Heteroscedasticity	77
(iii) Autocorrelation	82

3 Further Aspects of Least-squares Methods	94
3.0 Introduction	94
3.1- Least Squares Curve-fitting	95
3.1.0 Introduction	95
3.1.1 Gram-Schmidt Method: Computation	96
3.1.2 Examples	101
3.1.3 Confidence Intervals for Least-squares Coefficients	109
3.2 Trend Surface Analysis	116
3.2.0 Introduction	116
3.2.1 Trend Surface Model	117
(A) Specification Phase	119
(B) Data Point Distributions	120
(C) Computational Aspects	130
(D) Inferential Problems	142
3.2.2 Comparison of Trend Surfaces	147
3.2.3 Examples	150
3.3 Selection of Explanatory Variables in Multiple Regression	172
3.3.0 Introduction	172
3.3.1 The Selection Approach	173
3.3.2 The Combinational Approach	178
3.4 Use of Binary Explanatory Variables	193
3.4.0 Introduction	193
3.4.1 Method and Example	194
3.5 Nonlinear Least Squares Methods	198
3.5.0 Introduction	198
3.5.1 Methods	198
3.5.2 Example	207

PART THREE ANALYSIS OF INTERDEPENDENCE

4 Principal Components Analysis	215
4.0 Introduction	215
4.1 Mathematical Foundations	220
4.2 Computation	226
4.3 Example	232
4.4 Conclusions	239
5 Factor Analysis	240
5.0 Introduction	240
5.1 Definitions	241
5.2 Statistical Model	243
5.3 Estimation	245
5.3.0 Nonstatistical Factor Analysis	247
5.3.1 Statistical Factor Analysis	261

5.3.2	Image Analysis	272
5.4	Rotation of Factors	275
5.5	Factor Scores	294
5.6	Summary	304

PART FOUR CLASSIFICATION

6	Classification	309
6.0	Introduction	309
6.1	Cluster Analysis	310
6.1.0	Types of Cluster Analysis	310
6.1.1	Measurement of Similarity	311
6.1.2	Hierarchical Clustering Methods	316
6.1.3	Display of Hierarchical Relationships	321
6.1.4	Comparison of Hierarchical Clustering Strategies	324
6.1.5	Nonhierarchical Clustering Methods	327
6.2	Ordination	329
6.2.0	Principal Coordinates Analysis	330
6.2.1	Nonmetric Multidimensional Scaling	331
6.2.2	Nonlinear Mapping	339
6.2.3	Quadratic Loss Functions	340
6.3	Examples	341
6.3.0	Hierarchical Clustering	342
6.3.1	Nonhierarchical Cluster Analysis; Nonlinear Mapping	343
6.3.2	Nonmetric Multidimensional Scaling	351
6.3.3	Principal Coordinates Analysis	370
6.4	Computer Programs	371
7	Discriminant Analysis	420
7.0	Introduction	420
7.1	Two-group Discriminant Analysis	420
7.2	Multiple Discriminant Analysis	432
7.2.0	Multivariate Analysis of Variance (MANOVA)	432
7.2.1	Multiple Discriminant Analysis	437
7.3	Discriminant Analysis based on Binary Attributes	453
Appendix A	Fortran Subroutine Library	460
Bibliography		504
Author Index		521
Subject Index		527

Introduction

0.0 Exploratory Data Analysis

The success of modern science in arriving at an understanding of the way in which physical phenomena work and are interrelated is largely due to the examination of supposed causal relationships. In many research situations in pure science, cause and effect have been isolated by the study of relationships between pairs of variables. Experiments carried out under laboratory conditions have made it possible to isolate a few variables of interest, with the other controlling variables being held constant. Physical geography, however, is concerned with complex situations in which many variables are simultaneously related (or even related to their own past values). This results in the obscuring of causal links. Since laboratory studies of the kind mentioned above are generally not feasible, the complex of variables must be studied as a whole. This is due to the fact that changes in one variable may produce corresponding changes in other variables either directly or indirectly, making it very difficult to isolate pairs of strongly-related variables as is possible in many laboratory-oriented sciences.

Chorley and Kennedy (1971) have drawn attention to the advantages of a systems approach in physical geography. Such an approach allows the set of variables under consideration to be viewed as a whole, rather than as the sum of a number of simpler relationships. The behaviour of a complex interacting system of variables cannot easily or efficiently be described in terms of a number of bivariate relationships. In order to understand such systems, it is necessary to use multivariate analysis.

Two opposing views have been taken by users of multivariate methods. Some see the role of these methods as a purely confirmatory one, that is, providing estimates of the parameters of *a priori* models, and allowing significance tests of hypotheses relating to the values taken by these parameters. This attitude is perhaps more relevant to research in the pure sciences, where a firm basis of theory has been established. In other, less well-developed, disciplines multivariate

methods are commonly used in an exploratory manner, that is, to suggest, rather than to test, hypotheses. In fact, it could be argued that changes in the orientation of a discipline are more often due to developments in techniques than to changes in theory. New techniques allow the research worker to ask questions of a different, perhaps more complicated, nature. Such initial uses of new techniques are inevitably exploratory rather than confirmatory. (Consider, for example, the state of astronomy after the invention of the telescope.) It is interesting to note that the period 1956 to 1963, during which geography underwent a major internal reorganization, is usually known as the 'quantitative revolution' even though, as Burton (1963) and other writers have shown, the fundamental revolution was in the methodological outlook of geographers. Perhaps the availability of statistical techniques, coupled with the ability to use them, provoked geographers to think of alternative ways of explaining the phenomena and relationships they were scrutinising. However, the 'quantitative revolution' was not necessarily entirely beneficial. Geographers have perhaps been too ready to proclaim themselves to be scientists, and have, as a consequence, eagerly assumed what they thought was 'the' scientific method. Such eagerness in a developing subject can lead to premature regimentation and sterile debate at a time when trial and error methods should be used more freely. Consider a child of three. He does not learn to speak fluently and grammatically by adopting a 'scientific' method of learning; he listens and he experiments with new words and new verbal structures, rejecting some and accepting others on the basis of his success in communicating. In that way, he builds up a body of experience which — at a later stage — allows him to put together words and phrases on the basis of his theoretical knowledge of the language. Despite what St. Paul had to say about casting off childish ways, this analogy has much to commend it.

Exploratory data analysis is, then, a method of examining a set of data from various angles, and of piecing together information about the system of study that is revealed by each analysis. Such information may lead to a subsequent analysis that is more refined and possibly more revealing. The lack of a formal hypothesis may, in some cases, lead to error. However, the possibility of error is not in itself a sufficient reason to reject or to condemn exploratory methods. Knowledge is gained only if the research worker is willing to take the risk of making mistakes resulting in faulty conclusions. Tukey (1969), in a paper entitled *Analysing Data: Sanctification or Detective Work?* makes essentially the same point: '... to concentrate on confirmation, to the exclusion or submergence of exploration, is an obvious mistake. Where does new knowledge come from? How can an undetected criminal be put on trial? ... we learn by taking chances ... no conclusion or inference ever becomes knowledge without risk or error.'

Repetition is important in exploratory work. If it is not possible to take further samples from the population of interest, the given sample can be split into several subsamples, each of which should then be analysed independently. Only results that are consistent across samples should be considered for further investigation. Another aspect of exploratory data analysis that is worthy of comment is the status of uninterpretable results. Such results should not be discarded lightly. Quite often, it is the presence of anomalous results that reveals deficiencies in the form of the model (is the model correctly specified? have important explanatory variables been omitted, or irrelevant ones included?) As Digman (1966, p. 471) remarks: '... there are few things which will stimulate research and thinking as much as a stable but inscrutable phenomenon.' In this context, it could be said that learning comes from the interplay between the investigator, the technique, and the data.

It is the author's belief that such an interplay is not possible unless the investigator has a good basic knowledge of the techniques he is using, in addition to a firm general grasp of the system he is studying. Consequently, emphasis is placed in this book on the technical aspects of the specification and assumptions of the multivariate models described, and to computational and other practical difficulties that are often experienced in their application. This is the opposite approach to that favoured by many subject-oriented texts on quantitative methods, the authors of which appear to prefer to work from a case study to the details of the techniques. Since it is assumed that the reader will be actively engaged in research, and will be aware of problems in his own field that are amenable to a multivariate approach, extended case studies are not given. The examples that are included are not meant to be representative applications of the techniques — they are used to show the nature of the results to be expected from a given technique, and to provide test data for the computer programs that are listed in each chapter. References are given to selected applications of the individual techniques, but no comprehensive review of the literature of quantitative physical geography is provided. The reader is recommended to refer to the specialist literature in his own field for details of relevant uses of the various techniques described in this book.

0.1 Structure of this book

Readers are assumed to have some familiarity with elementary statistical concepts as laid out by, among others, Agterberg (1974), Kmenta (1971) and Yamane (1964). Since a knowledge of matrix algebra is essential if the material in later chapters is to be intelligible, Part 1 of this book consists of a survey of matrix concepts and operations. No prior knowledge of matrices is assumed. Hand-calculator

methods of carrying out the operations described in this part are not given; in all but the most trivial cases the computations will not be performed by hand, and it seems unnecessary to labour the reader with details of outmoded algorithms which, for the most part, are unsuitable for computer implementation. Fortran subroutines to carry out the operations that are mentioned in the text are provided in Appendix A. This Appendix also contains listings of general purpose routines. Attention is given in the closing section of Chapter 1 to the question of computational accuracy in multivariate analysis. This important topic has been generally ignored, and not only in the geographical literature, although it should be self-evident that a computational algorithm should not distort the data. Many readers will, no doubt, be familiar with the phrase 'garbage in, garbage out' but it is true that good data are sometimes transformed into garbage by an inefficient or unstable algorithm. Part 1 closes with an appendix containing a summary of matrix algebra.

The second part of the book consists of a survey of least-squares methods. Chapter 2 deals with the classical ordinary least-squares model and with the strategy to be followed whenever the basic assumptions of the model do not hold. Despite the popularity of regression methods over the last 20 years, many users appear to be unaware of the nature of these assumptions, and few applications of alternative methods of estimation have appeared in the literature of physical geography. In an exploratory analysis it will not normally be known whether or not the disturbance terms in a multiple regression analysis are heteroscedastic or autocorrelated; tests on the residuals from an ordinary least-squares regression provide the opportunity to examine hypotheses about the residuals and allow the investigator to modify the specification of the model either by altering its form or by introducing explanatory variables that were not previously considered. Also, alternative forms of estimation can be employed if such modifications do not produce a model of the classical form. For example, generalized least-squares procedures can be used in the presence of autocorrelated or heteroscedastic disturbances, or ridge regression methods if the explanatory variables are highly correlated.

Extensions of the classical model are dealt with in Chapter 3. Methods of fitting polynomials in one and two explanatory variables (trend surface analysis) are introduced, and the computational and logical problems associated with these techniques are discussed. An important application of regression analysis in exploratory studies is concerned with the selection of important explanatory variables. Two selection methods are discussed — the now traditional stepwise approach and the combinatorial method, which has been made feasible by improvements in the size and speed of modern digital computers. The last two sections in Chapter 3 are concerned with the use of binary

explanatory variables and the fitting of nonlinear least-squares models.

Chapters 4 and 5, which form Part 3 of the book, contain a review of methods of analysing the dependence structures within a system of interrelated variables. Principal components analysis is the topic of Chapter 4, while an account of the various methods of factor analysis is given in Chapter 5. Some confusion over the aims of these two techniques is evident from even a casual inspection of the geographical literature. In the author's opinion, principal components analysis is a mathematical transformation of a data matrix to produce a new matrix of scores that has the property that successive orthogonal columns (representing principal components) account for a maximum of the variance or information remaining after the amounts accounted for by preceding columns have been removed. Principal components are thus defined in terms of variance. The principal components score matrix contains the same information as the data matrix, but that information is expressed in a compressed, more economical form. Concepts such as rotation and communality are extraneous to principal components analysis. Factor analysis, on the other hand, is concerned with the structure of the correlations between the variables, and is therefore a suitable technique to use if the pattern of relationships within a system is being scrutinized. Factors are not defined in terms of variance; they are defined in accordance with a hypothesis such as that of simple structure. The methods of factor analysis have recently undergone a considerable revision, and more modern methods — such as those described in this book — have done much to alleviate the criticisms of the technique that have been voiced over the last few years. In particular, the principal axis method has been seen to be a computational expedient rather than a true factoring method.

The fourth and final part of the book is devoted to methods of classification and discrimination. All developing sciences are pre-occupied with the classification of the phenomena with which they deal, in order to reduce the number of concepts necessary to understand a particular system. No single classification can be expected to serve all purposes, hence the research worker must specify his own requirements before adopting an existing classification or constructing a new one. Many methods of numerical classification have been described in the literature of various disciplines. Some of the more useful of these are discussed in Chapter 6. Once a suitable classification has been arrived at, it may be necessary to allocate previously unclassified items or observations to their most likely group while minimizing the possibility of error. Chapter 7 is a survey of techniques of discriminant analysis including two-group and multiple-group discriminant analysis as well as a summary of more recent methods of pattern recognition that are suitable for use when the variables are dichotomous rather than continuous in nature.

0.2 Computer Programs

Chapters 2 to 7 contain listings of Fortran computer programs to carry out the calculations described in the text. All have been written specifically for this book. Although some of the techniques have previously been programmed (for example, they may be available in one of the standard packages, such as BMD or SPSS) the versions listed here are more flexible, allowing the user to choose between various alternative methods of input and output, for example. Furthermore, because the source listings are provided, and details of the computations are discussed in the text, the reader should be able to modify the programs to suit his own particular requirements. Where alternative estimation procedures are available, for example in factor analysis, individual subroutines are provided. To help readers program the algorithms for which no Fortran listing is included, a library of matrix handling and general purpose subroutines is given in Appendix A. Machine-dependent aspects of Fortran have been avoided as far as possible, and many of the programs have been tested on a small computer (PDP 11), a medium size computer (ICL 1906A) and a large computer (CDC 7600). All of the few machine-dependent instructions are indicated by comments within the listings. The matrix handling routines in Appendix A are based in part on the Algol 60 procedures of Wilkinson and Reinsch (eds.) (1971), to which reference should be made for more detailed computational and algorithmic details.

Part One

Essentials of Matrix Algebra

Two aspects of matrices are dealt with in this Part; (i) the concepts and notation of matrix algebra, and (ii) computational methods for matrices. Matrix algebra provides a concise, abbreviated notation which is well suited to the expression of multivariate relationships. The material covered in the remainder of this book requires a basic knowledge of matrices, including addition, subtraction, multiplication and inversion of matrices; the solution of systems of linear simultaneous equations; the derivation of the eigenvalues and eigenvectors of matrices, and the choice of a computational algorithm to carry out these operations. Geographers, like many other users of multivariate analysis, have tended to ignore the numerical analysis aspects of the topic, possibly due to a belief that the computer can carry out any given operation to the required degree of accuracy, or else to a belief that computational errors will always be of less importance than data measurement errors. Neither of these suppositions is justified, and the purpose of this Part is to show both the utility of matrix methods and the safeguards that are necessary in their use.