MATHEMATICAL TOOLS FOR APPLIED MULTIVARIATE ANALYSIS

Paul E. Green

J. Douglas Carroll

MATHEMATICAL TOOLS FOR APPLIED MULTIVARIATE ANALYSIS

Paul E. Green

Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania

with contributions by

J. Douglas Carroll

Bell Laboratories
Murray Hill, New Jersey

ACADEMIC PRESS New York San Francisco London 1976

A Subsidiary of Harcourt Brace Jovanovich, Publishers

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

Library of Congress Cataloging in Publication Data

Green, Paul E
Mathematical tools for applied multivariate analysis.

Bibliography: p.
1. Multivariate analysis. I. Title.
QA278.G73 519.5'3 75-36647
ISBN 0-12-297550-2

PRINTED IN THE UNITED STATES OF AMERICA

Preface

The student willing to learn something about multivariate analysis will find no dearth of textbooks and monographs on the subject. From introductory to advanced, theoretical to applied, general to specific, the field has been well covered.

However, most of these books assume certain mathematical prerequisites—typically matrix algebra and introductory calculus. Single-chapter reviews of the topics are usually provided but, in turn, presuppose a fair amount of advance preparation. What appears to be needed for the student who has received less exposure is a somewhat more elementary and leisurely approach to developing the necessary mathematical foundations of applied multivariate analysis.

The present book has been prepared to help students with those aspects of transformational geometry, matrix algebra, and the calculus that seem most relevant for the study of multivariate analysis. Since the author's interest is in applications, both the material selected for inclusion and the point of view from which it is presented reflect that orientation.

The book has been prepared for students who have either taken no matrix algebra at all or, if they have, need a refresher program that is between a full-fledged matrix algebra course and the highly condensed review chapter that is often found in multivariate textbooks. The book can serve as a textbook for courses long enough to permit coverage of precursory mathematical material or as a supplement to general textbooks on multivariate analysis.

The title was chosen rather carefully and helps demarcate what the book is not as much as what it is. First, those aspects of linear algebra, geometry, and the calculus that are covered here are treated from a pragmatic viewpoint—as tools for helping the applications researcher in the behavioral and business disciplines. In particular, there are virtually no formal proofs. In some cases outlines of proofs have been sketched, but usually small numerical examples of the various concepts are presented. This decision has been deliberate and it is the author's hope that the instructor will complement the material with more formal presentations that reflect his interests and perceptions of the technical backgrounds of his students.

The book consists of six chapters and two appendixes. Chapter 1 introduces the topic of multivariate analysis and presents three small problems in multiple regression, principal components analysis, and multiple discriminant analysis to motivate the mathematics that subsequent chapters are designed to supply. Chapter 2 presents a fairly standard treatment of the mechanics of matrix algebra including definitions and operations on vectors, matrices, and determinants. Chapter 3 goes through much of this same material but from a geometrically oriented viewpoint. Each of the main ideas in matrix algebra is illustrated geometrically and numerically (as well as algebraically).

Chapters 4 and 5 deal with the central topics of linear transformations and eigenstructures that are essential to the understanding of multivariate techniques. In Chapter 4, the theme of Chapter 3 receives additional attention as various matrix transformations are illustrated geometrically. This same (geometric) orientation is continued in Chapter 5 as eigenstructures and quadratic forms are described conceptually and illustrated numerically.

Chapter 6 completes the cycle by returning to the three applied problems presented in Chapter 1. These problems are solved by means of the techniques developed in Chapters 2–5, and the book concludes with a further discussion of the geometric aspects of linear transformations.

Appendix A presents supporting material from the calculus for deriving various matrix equations used in the book. Appendix B provides a basic discussion on solving sets of linear equations and includes an introduction to generalized inverses. Numerical exercises appear at the end of each chapter and represent an integral part of the text. With the student's interest in mind, solutions to all numerical problems are provided. (After all, it was those *even*-numbered exercises that used to give us all the trouble!) The student is urged to work through these exercises for purposes of conceptual as well as numerical reinforcement.

Completion of the book should provide both a technical base for tackling most applications-oriented multivariate texts and, more importantly, a geometric perspective for aiding one's intuitive grasp of multivariate methods. In short, this book has been written for the student in the behavioral and administrative sciences—not the statistician or mathematician. If it can help illuminate some of the material in current multivariate textbooks that are designed for this type of reader, the author's objective will have been well satisfied.

Acknowledgments

Many people helped bring this book to fruition. Literally dozens of masters and doctoral students provided critical reviews of one or more chapters from the most relevant perspective of all—their's. Those deserving special thanks are Ishmael Akaah, Alain Blancbrude, Frank Deleo, J. A. English, Pascal Lang, and Gunter Seidel. Professor David K. Hildebrand, University of Pennsylvania, prepared a thorough and useful critique of the full manuscript. Helpful comments were also received from Professor Joel Huber, Purdue University.

Production of the book was aided immeasurably by the competent and cheerful help of Mrs. Joan Leary, who not only typed drafts and redrafts of a difficult manuscript, but managed to do the extensive art work as well. The editorial staff of Academic Press also deserves thanks for their production efforts and general cooperative spirit.

The author's biggest debt of gratitude is to J. Douglas Carroll of Bell Laboratories. His imprint on the book's organization and exposition goes well beyond the role of reviewer. While the specific words in the book are mine, its general intent and orientation are fully shared with Dr. Carroll. (However, he is to be held blameless for any of the words that did not come out quite right.)

P. E. G.

Contents

CHAPTER 1

The Nature of Multivariate Data Analysis

1.1 INTRODUCTION

Stripped to their mathematical essentials, multivariate methods represent a blending of concepts from matrix algebra, geometry, the calculus, and statistics. In function, as well as in structure, multivariate techniques form a unified set of procedures that can be organized around a relatively few prototypical problems. However, in scope and variety of application, multivariate tools span all of the sciences.

This book is concerned with the mathematical foundations of the subject, particularly those aspects of matrix algebra and geometry that can help illuminate the structure of multivariate methods. While behavioral and administrative applications are stressed, this emphasis reflects the background of the author more than any belief about special advantages that might accrue from applications in these particular fields.

Multivariate techniques are useful for:

1. discovering regularities in the behavior of two or more variables;
2. testing alternative models of association between two or more variables, including the determination of whether and how two or more groups (or other entities) differ in their "multivariate profiles."

The former pursuit can be regarded as exploratory research and the latter as confirmatory research. While this view may seem a bit too pat, multivariate analysis *is* concerned with both the discovery and testing of patterns in associative data.

The principal aim of this chapter is to present motivational material for subsequent development of the requisite mathematical tools. We start the chapter off on a somewhat philosophical note about the value of multivariate analysis in scientific research generally. Some of the major characteristics of multivariate methods are introduced at this point, and specific techniques are briefly described in terms of these characteristics.

Application of multivariate techniques is by no means confined to a single discipline. In order to show the diversity of fields in which the methods have been applied, a number of examples drawn from the behavioral and administrative sciences are briefly described. Comments are also made on the trends that are taking place in multivariate analysis itself and the implications of these developments for future application of the methodology.

We next turn to a description of three small, interrelated problems that call for multivariate analysis. Each problem is described in terms of a common, miniature data bank with integer-valued numbers. As simple as the problems are, it turns out that developing the apparatus necessary to solve them covers most of the mathematical concepts in multivariate analysis that constitute the rest of the book.

1.2 MULTIVARIATE METHODS IN RESEARCH

It is difficult to imagine any type of scientific inquiry that does not involve the recording of observations on one or more types of objects. The objects may be things, people, natural or man-made events. The selected objects–white rats, model airplanes, biopsy slides, x-ray pictures, patterns of response to complex stimulus situations, ability tests, brand selection behavior, corporate financial activities–vary with the investigator's discipline. The process by which he codifies the observations does not.

Whatever their nature, the objects themselves are never measured in total. Rather, what is recorded are observations dealing with *characteristics* of the objects, such as weight, wind velocity, cell diameter, location of a shadow on the lung, speed or latency of response, number of correctly answered questions, specific brand chosen, previous year's sales, and so on. It is often the case that two or more characteristics (e.g., weight, length, and heartbeat) will be measured at the same time on each object being studied. Furthermore, it would not be unusual to find that the measured characteristics were associated in some way; that is, values taken on by one variable are frequently related to values taken on by another variable.

As a set of statistical techniques, multivariate data analysis is strategically neutral. Techniques can be used for many purposes in the behavioral and administrative sciences–ranging from the analysis of data obtained from rigidly controlled experiments to teasing out relationships assumed to be present in a large mass of survey-type data. What can be said is that multivariate analysis is concerned with *association among multiple variates* (i.e., many variables).[1]

Raymond Cattell (1966) has put the matter well. Historically, empirical work in the behavioral sciences–more specifically, experimental psychology–has reflected two principal traditions: (a) the manipulative, typically bivariate approach of the researcher viewed as controller and (b) the nonmanipulative, typically multivariate approach of the researcher viewed as observer.

Cattell points out three characteristics that serve to distinguish these forms of strategic inquiry:

1. bivariate versus multivariate in the type of data collected,
2. manipulative versus *noninterfering* in the degree of control exercised by the researcher,
3. simultaneous versus temporarily successive in the time sequence in which observations are recorded.

[1] Analysis of bivariate data can, of course, be viewed as a special case of multivariate analysis. However, in this book our discussion will emphasize association among more than two variables. One additional point–some multivariate statisticians restrict the term *multivariate* to cases involving more than a single criterion variable. Here, we take a broader view that includes multiple regression and its various extensions as part of the subject matter of multivariate analysis.

In recent years, bivariate analysis and more rigid forms of controlled inquiry have given way to experiments and observational studies dealing with a comparatively large number of variables, not all of which may be under the researcher's control. However, if one takes a broad enough view of multivariate data analysis, one that includes bivariate analysis as a special case, then the concepts and techniques of this methodology can be useful for either stereotype. Indeed, Cattell's definition of an experiment as:

> . . .A recording of observations, quantitative or qualitative, made by defined operations under defined conditions, and designed to permit non-subjective evaluation of the existence or magnitude of relations in the data. It aims to fit these relations to parsimonious models, in a process of hypothesis creation or hypothesis checking, at least two alternatives being logically possible in checking this fit. . . . (p. 9)

says quite a bit about the purview of multivariate analysis. That is, the process of scientific inquiry should embrace the search for naturalistic regularities in phenomena as well as their incorporation into models for subsequent testing under changed conditions. And in this book we shall be as much, if not more so, interested in using multivariate analysis to aid the process of discovery (hypothesis creation) as to aid the process of confirmation (hypothesis testing).

The heart of any multivariate analysis consists of the data matrix, or in some cases, matrices.[2] The data matrix is a rectangular array of numerical entries whose informational content is to be summarized and portrayed in some way. For example, in univariate statistics the computation of the mean and standard deviation of a single column of numbers is often done simply because we are unable to comprehend the meaning of the entire column of values. In so doing we often (willingly) forego the full information provided by the data in order to understand some of its basic characteristics, such as central tendency and dispersion. Similarly, in multivariate analysis we often use various summary measures–means, variances, covariances–of the raw data. Much of multivariate analysis is concerned with placing in relief certain aspects of the association among variables at the expense of suppressing less important details.

In virtually all applied studies we are concerned with variation in some characteristic, be it travel time of a white rat in a maze or the daily sales fluctuations of a retail store. Obviously, if there is no variation in the characteristic(s) under study, there is little need for statistical methods.

In multivariate analysis we are often interested in accounting for the variation in one variable or group of variables in terms of *covariation* with other variables. When we analyze associative data, we hope to "explain" variation according to one or more of the following points of view:

1. determination of the nature and degree of association between a set of *criterion* variables and a set of *predictor* variables, often called "dependent" and "independent" variables, respectively;

2. finding a function or formula by which we can estimate values of the criterion variable(s) from values of the predictor variable(s)–this is usually called the *regression* problem;

[2] Much of this section is drawn from Green and Tull (1975).

3. assaying the statistical "confidence" in the results of either or both of the above activities, via tests of statistical significance, placing confidence intervals on parameter estimates, or other ways.

In some cases of interest, however, we have no prior basis for distinguishing between criterion and predictor variables. We may still be interested in their interdependence as a whole and the possibility of summarizing information provided by this interdependence in terms of other variables, often taken to be linear composites of the original ones.

1.3 A CLASSIFICATION OF TECHNIQUES FOR ANALYZING ASSOCIATIVE DATA

The field of associative data analysis is vast; hence it seems useful to enumerate various descriptors by which the field can be classified. The key notion underlying the classification of multivariate methods is the *data matrix.* A conceptual illustration is shown in Table 1.1. We note that the table consists of a set of objects (the m rows) and a set of measurements on those objects (the n columns). Cell entries represent the value X_{ij} of object i on variable j. The objects are any kind of entity with characteristics capable of being measured. The variables are characteristics of the objects and serve to define the objects in any specific study. The cell values represent the state of object i with respect to variable j. Cell values may consist of nominal, ordinal, interval, or ratio-scaled measurements, or various combinations of these, as we go across columns.

By a nominal scale we mean categorical data where the only thing we know about the object is that it falls into one of a set of mutually exclusive and collectively exhaustive categories that have no necessary order vis à vis one another. Ordinal data are ranked data where all we know is that one object i has more, less, or the same amount of some variable j than some other object i'. Interval scale data enable us to say how much more one object has than another of some variable j (i.e., intervals between scale values are meaningful). Ratio scale data enable us to define a natural origin (e.g., a case in which

TABLE 1.1

Illustrative Data Matrix

	Variables							
Objects	1	2	3		j		n
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1j}	$\cdots$	x_{1n}
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2j}	$\cdots$	x_{2n}
3	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3j}	$\cdots$	x_{3n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_{i1}	x_{i2}	x_{i3}		x_{ij}		x_{in}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
m	x_{m1}	x_{m2}	x_{m3}	$\cdots$	x_{mj}	$\cdots$	x_{mn}

object *i* has zero amount of variable *j*), and ratios of scale values are meaningful. Each higher scale type subsumes the properties of those below it. For example, ratio scales possess all the properties of nominal, ordinal, and interval scales, in addition to a natural origin.

There are many descriptors by which we can characterize methods for analyzing associative data.[3] The following represent the more common bases by which the activity can be classified:

1. purpose of the study and the types of assertions desired by the researcher—what kinds of statements does he wish to make about the data or about the universe from which the data were drawn?
2. focus of research emphasis—statements regarding the objects (i.e., the whole profile or "bundle" of variables), specific variables, or both;
3. nature of his prior judgments as to how the data matrix should be partitioned in terms of the type and number of subsets of variables;
4. number of variables in each of the partitioned subsets;
5. type of association under study—linear in the parameters, transformable to linear, or "inherently" nonlinear in the parameters;
6. scales by which variables are measured—nominal, ordinal, interval, ratio, mixed.

All of these descriptors relate to certain decisions required of the researcher. Suppose he is interested in studying certain descriptive relationships among variables. If so, he must make decisions about how he wants to partition the set of columns (see Table 1.1) into subsets. Often he will call one subset "criterion" variables and the other subset "predictor" variables.[4] He must also decide, however, on the number of variables to include in each subset and on what type of functional relationship is to hold among the parameters in his statistical model.

Most decisions about associative data analysis are based on the researcher's "private model" of how the variables are related and what features are useful for study.[5] His choice of various "public models" for analysis—multiple regression, discriminant analysis, etc.—is predicated on his prior knowledge of the characteristics of the statistical universe from which the data were obtained and his knowledge of the assumption structure of each candidate technique.

1.3.1 Researcher's Objectives and Predictive Statements

We have already commented that the researcher may be interested in (a) measuring the nature and degree of association between two or more variables; (b) predicting the values of one or more criterion variables from values of one or more predictor variables; or (c)

[3] An excellent classification, based on a subset of the descriptors shown here, has been provided by M. M. Tatsuoka and D. V. Tiedeman (1963).

[4] As Horst (1961) has shown, relationships need not be restricted to two sets.

[5] To some extent this is true even of the scales along which the data are measured. The researcher may wish to "downgrade" data originally expressed on interval scales to ordered categories, if he feels that the quality of the data does not warrant the "strength" of scale in which it is originally expressed. In other cases he may "upgrade" data in order to use some statistical technique that assumes a type of measurement that is absent originally.

assessing the statistical reliability of an association between two or more variables. In a specific study all three objectives may be pursued. In using other techniques (i.e., those dealing mainly with factor and cluster analysis), the researcher may merely wish to portray association in a more parsimonious way without attempting to make specific predictions or inferential statements.

1.3.2 Focus of Research Interest

Some multivariate techniques (e.g., multiple regression) focus on association among variables; objects are treated only as replications. Other techniques (e.g., cluster analysis) focus on association among objects; information about specific variables is usually, although not necessarily, suppressed. In still other instances one may wish to examine interrelationships among variables, objects, and object–variable combinations, as well.

1.3.3 Nature of Assumed Prior Judgments or Presuppositions

In many cases the investigator is able to partition the data matrix into subsets of columns (or rows) on the basis of prior judgment. For example, suppose the first column of Table 1.1 is average weekly consumption of coffee by households, and the other columns consist of various demographic measurements of the m households. The analyst may wish to predict average weekly consumption of coffee from some linear composite of the $n - 1$ remaining variables. If so, he has used his presuppositions regarding how the dependence is to be described and, in this instance, might employ multiple regression.

In most cases the number of subsets developed from the data matrix partitioning will be two, usually labeled as criterion and predictor variable subsets. However, techniques have been designed to summarize association in cases involving more than two subsets of data.

Finally, we may have no reasonable basis for partitioning the data matrix into criterion or predictor variables. Our purpose here may be merely to group objects into "similar" subsets, based on their correspondence over the whole profile of variables. Alternatively, we may wish to portray the columns of the data matrix in terms of a smaller number of variables, such as linear combinations of the original set, that retain most of the information in the original data matrix. Cluster analysis and factor analysis, respectively, are useful techniques for these purposes.

1.3.4 Number of Variables in Partitioned Subsets

Clearly, the term "association" implies at least two characteristics—for example, a single criterion and a single predictor variable, usually referred to as bivariate data. In other cases involving two subsets of variables, we may wish to study association between a single criterion and more than one predictor. Or we may wish to study association between composites of several criterion variables and composites of several predictor variables. Finally we may want to study the relationship between several criterion variables and a single predictor variable.

Of course, we may elect not to divide the variables at all into two or more subsets, as would be the case in factor analysis. Furthermore, if we do elect to partition the matrix and end up with two or more variables in a particular subset, what we are usually concerned with are various *linear composites* of the variables in that subset and each composite's association with other variables.

1.3.5 Type of Association

Most of the models of multivariate analysis emphasize linear relationships among the variables. The assumption of linearity, in the parameters, is not nearly so restrictive as it may seem.[6] First, various preliminary transformations (e.g., square root, logarithmic) of the data are possible in order to achieve linearity in the parameters.[7] Second, the use of "dummy" variables, coded, for example, as elementary polynomial functions of the "real" variables, or indicating category membership by patterns of zeroes and ones, will enable us to handle certain types of nonlinear relationships within the framework of a linear model. Third, a linear model is often a good approximation to a nonlinear one, at least over restricted ranges of the variables in question.

1.3.6 Types of Scales

Returning to the data matrix of Table 1.1, we now are concerned with the scales by which the characteristics are represented. Since all of the multivariate statistical techniques to be discussed in this book require no stronger form of measurement than an interval scale, we shall usually be interested in the following types: (a) nominal, (b) ordinal, and (c) interval. In terms of nominal scaling we shall find it useful to distinguish between dichotomies and (unordered) polytomies, the latter categorization involving more than two classes.

This distinction is important for three reasons. First, many of the statistical techniques for analyzing associative data are amenable to binary-coded (zero–one) variables but *not* to polytomies. Second, any polytomy can be recoded as a set of dichotomous "dummy" variables; we shall describe how this recoding is done in the next section. Third, when we discuss geometrical representations of variables and/or objects, dichotomous variables can be handled within the same general framework as interval-scaled variables.

Finally, mention should be made of cases in which the analyst must contend with *mixed* scales in the criterion subset, predictor subset, or both. Many multivariate techniques—if not modified for this type of application—lead to rather dubious results under such circumstances.

[6] By linear in the parameters is meant that the b_j's in the expression $y = b_1x_1 + b_2x_2 + \cdots + b_nx_n$ are each of the first degree. Similarly, $z = b_1x_1^2 + \cdots + b_nx_n^{n+1}$ is still linear in the parameters since each b_j continues to be of the first degree even though x_j is not.

[7] For example, the complicated expression $y = ax^{b}e^{cx}$ (with both a, $x > 0$) can be "linearized" as $\ln y = \ln a + b \ln x + cx$ and, as shown by Hoerl (1954), is quite flexible in approximating many diverse types of curves. On the other hand, the function $y = 1/(a + b^{-cx})$ is inherently nonlinear in the parameters and cannot be "linearized" by transformation.

1.4 ORGANIZING THE TECHNIQUES

In most textbooks on multivariate analysis, three of the preceding characteristics are often used as primary bases for technique organization:

1. whether one's principal focus is on the objects or on the variables of the data matrix;
2. whether the data matrix is partitioned into criterion and predictor subsets, and the number of variables in each;
3. whether the cell values represent nominal, ordinal, or interval scale measurements.

This schema results in four major subdivisions of interest:

1. *single criterion, multiple predictor association,* including multiple regression analysis of variance and covariance, and two-group discriminant analysis;
2. *multiple criterion, multiple predictor association,* including canonical correlation, multivariate analysis of variance and covariance, multiple discriminant analysis;
3. *analysis of variable interdependence,* including factor analysis, multidimensional scaling, and other types of dimension-reducing methods;
4. *analysis of interobject similarity,* including cluster analysis and other types of object-grouping procedures.

The first two categories involve dependence structures where the data matrix *is* partitioned into criterion and predictor subsets; in both cases interest is focused on the variables. The last two categories are concerned with interdependence—either focusing on variables or on objects. Within each of the four categories, various techniques are differentiated in terms of the type of scale assumed.

1.4.1 Scale Types

Traditionally, multivariate methods have emphasized two types of variables:

1. more or less continuous variables, that is, interval-scaled (or ratio-scaled) measurements;
2. binary-valued variables, coded zero or one.

The reader is no doubt already familiar with variables like length, weight, and height that can vary more or less continuously over some range of interest.

Natural dichotomies such as sex, male or female, or marital status, single or married, are also familiar. What is perhaps not as well known is that any (unordered) polytomy, consisting of three or more mutually exclusive and collectively exhaustive categories, can be recoded into dummy variables that are typically coded as one or zero. To illustrate, a person's occupation, classified into five categories, could be coded as:

Category	Dummy variable 1	2	3	4
Professional	1	0	0	0
Clerical	0	1	0	0
Skilled laborer	0	0	1	0
Unskilled laborer	0	0	0	1
Other	0	0	0	0