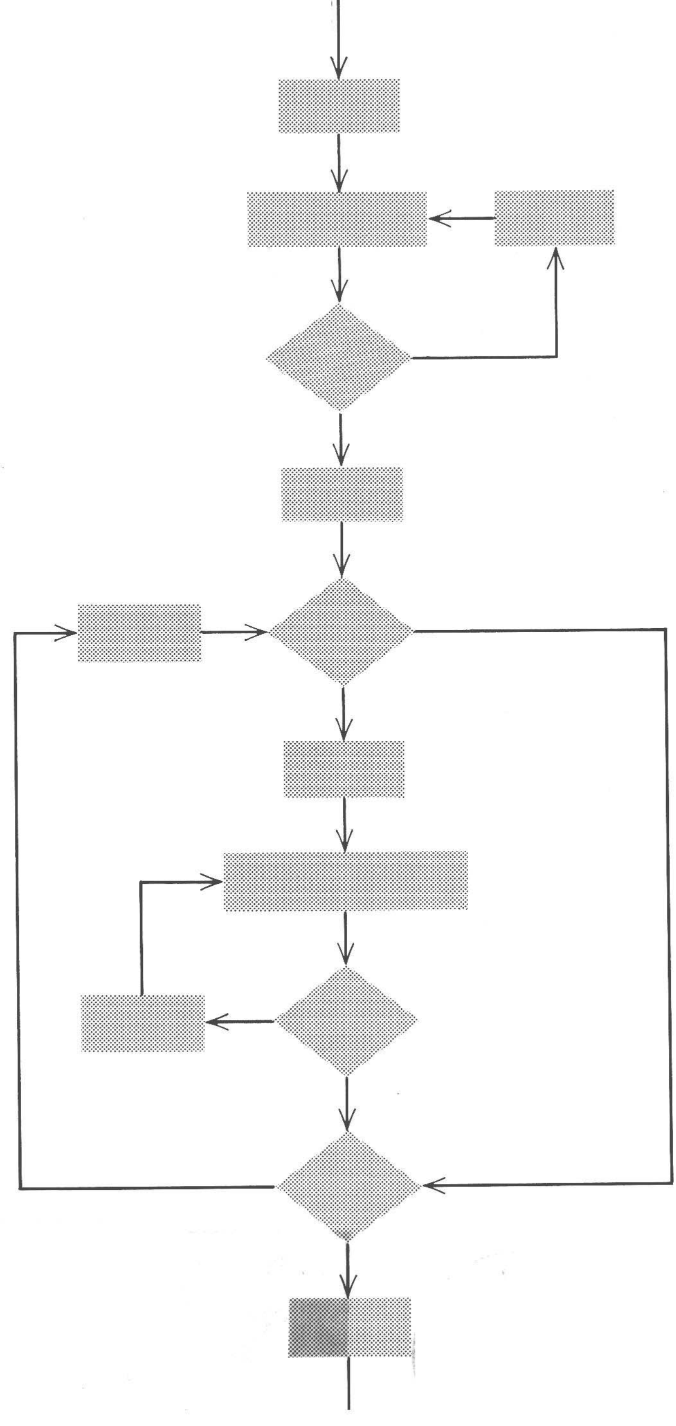


for

DIGITAL COMPUTERS

VOLUME III

of MATHEMATICAL METHODS FOR DIGITAL COMPUTERS

To the Memory of

HARRY HARMAN and LOUIS POTSDAMER

Library of Congress Cataloging in Publication Data (Revised)

Ralston, Anthony, ed.
Mathematical methods for digital computers.

Vol. 3 also has special title: Statistical methods for digital computers.
Vol. 3: Edited by K. Enslein, A. Ralston, and H. S. Wilf.
Includes bibliographies.
1. Mathematics—Data processing. I. Wilf, Herbert S., 1931– joint ed. II. Enslein, Kurt, joint ed. III. Title. IV. Title: Statistical methods for digital computers.

QA39.R26 519.5′3 60-6509
ISBN 0-471-70690-6 (v. 3)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

STATISTICAL METHODS *for* DIGITAL COMPUTERS

VOLUME III

Edited by

KURT ENSLEIN
Genesee Computer Center
Rochester, New York

ANTHONY RALSTON, Ph.D.
Professor of Mathematics and Director of Computer Center
State University of New York at Buffalo

HERBERT S. WILF, Ph.D.
Professor of Mathematics
The University of Pennsylvania

A Wiley-Interscience Publication
New York · London · Sydney · Toronto John Wiley & Sons, Inc.

STATISTICAL METHODS

PREFACE

Statistical computations account for a major portion of the jobs at any scientifically oriented computing facility. Volume I of this series did have four articles on statistical computation, but these are obsolete or very unsophisticated by current standards, and Volume II contained no articles in this area. The original editors (AR and HSW) felt, therefore, that it was surely time to devote a complete book to this most important area. The third editor (KE), who is an expert in the area of statistical computation with broad practical experience in this area, was invited to join in the production of this volume.

Our philosophy has been to include articles on as many of the most important and widely used techniques of statistical computation as possible in a book of approximately the same size as the previous two. We have tried to retain the main strengths of the previous volumes, in particular, by including essentially the same topics in each article and by keeping the format of each article as consistent as possible.

As in Volume II, we have included actual programs when the length of relevant programs is not excessively long. All such programs are in Fortran, still the language of statistical computation in an overwhelming majority of cases. Chapter 1 contains a guide to programs for all topics covered in the book.

To obtain full value from each chapter the reader should have at least some experience with elementary statistics, as well as a fair level of mathematical maturity. Others, however, will still benefit from using this book as a compendium of calculation procedures, flow charts, programs, and references.

Our contributors are outstanding research workers who have made important discoveries in the fields they discuss, and they have fully availed themselves of their opportunities for depth of discussion. We express our appreciation to them here, as well as the hope that the interested reader will benefit from their expertise, as we did.

KURT ENSLEIN
ANTHONY RALSTON
HERBERT S. WILF

Rochester, New York
Buffalo, New York
Philadelphia, Pennsylvania
April 1976

PART I

CONTENTS

Introduction

1. PURPOSE

In Volume 1 of *Mathematical Methods for Digital Computers* there were four chapters on statistical computation—Multiple Regression Analysis, Factor Analysis, Autocorrelation, and Spectral Analysis and Analysis of Variance. In 1960, when Volume 1 was published, that was a fairly representative coverage of the field of statistical computation. The last 15 years, however, have witnessed major growth in this area, and many new and valuable techniques of statistical computation are now in standard use. The 15 chapters of this volume are a sampling of the most important and most useful of these, but only a sampling, since no one volume could hope to be comprehensive because of the great scope of statistical computation today. Some areas of importance not covered by this volume are, however, mentioned briefly in Chapter 1.

2. FORMAT OF THE CHAPTERS

As in the previous two volumes of this series, the editors have tried to make the format as consistent as possible in as many chapters as possible. In this volume Chapter 1 does not follow the standard format because of its survey character and neither does Chapter 2 because its content does not lend itself well to the standard format. Chapters 3 through 15, however, all contain, with minor exceptions, the following materials:

Function

What is the function of the method that is about to be described? The contributor has endeavored to give a concise and accurate formulation of the particular problem which is considered in his chapter, sometimes in the context of prior art.

Mathematical Discussion

Insofar as the scope of the chapter extends, the contributor has tried to give a complete mathematical description of the problem which is to be solved and the method or methods by which he proposes to solve it. Here one may find relevant mathematical theorems with proofs or references to proofs, an error analysis if applicable and available including a discussion of the circumstances under which the technique may be expected to perform well or poorly, comparisons with other available techniques, and citations of the relevant literature.

Summary of the Calculation Procedure

All too often when a statistical discussion is interlaced with the derivation of a method the reader may finish quite unaware of what precise sequence of steps are required to carry out the solution of the problem. In this section, therefore, the reader will find the method previously derived stated in "recipe" form; that is, first do this, then this, and so on.

Flowchart

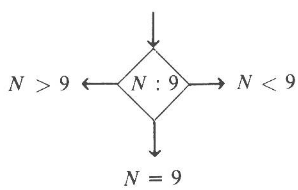
The flowcharts are as detailed as the generality of the material covered and the limitations of space allow.

The boxes in the flowcharts have been restricted to what is considered a bare minimum of different types. Flowcharts begin with a box labeled START and end with a box labeled STOP. The other types of boxes used are:

1. The assertion box. This box asserts that the operations contained within it are executed at this time.

$$\boxed{a + b \to c}$$

2. The test box.



This box has one input line and three possible output lines. When three output terminals are used one is chosen for greater than, less than, or equality output. When only two output terminals are used these indicate whether the answer to the question contained in the box is affirmative or negative.

3. The remote connector

This circle indicates that logical control is transferred to the point where another such circle appears with the same number in it.

Description of the Flowchart

A box-by-box description of the flowchart is given to aid the reader in following it.

Program and Subroutines

Due to space limitations, only crucial Fortran IV subroutines are provided. In any case, standard subroutines that would be required by a program are listed. Chapter 1 also provides the name of organizations from which the programs can be obtained.

Sample Problem

A representative sample problem is worked through to give the reader a feeling for the behavior of the process in practice.

References

The references cited in the text are listed. References are indicated in the text by numbers enclosed in square brackets.

A Guided Tour Through the Statistical Computing Garden

1

Kurt Enslein
Genesee Computer Center

1. INTRODUCTION

Whereas the previous two volumes in this series have dealt mainly with *mathematical* methods for digital computers, and only occasionally with *statistical* methods for digital computers, this volume deals almost exclusively with the latter. The great variety of statistical computing means that any work such as this cannot be totally comprehensive. The editors were, therefore, faced with difficult choices as to which methods to include and which to exclude. Those methods included are the ones we felt were of outstanding significance both because of their importance as statistical techniques and because of their widespread use on computers.

In this chapter we review the present state of the art of statistical computing, and more particularly *multivariate* statistical computing with particular relevance to the chapters in this volume. For an excellent review of recent trends of development in multivariate analysis we refer the reader to the comprehensive article by C. R. Rao [1].

The present author considered providing descriptive text elaborating on each of the techniques covered in this volume as well as descriptions of some that are not covered. However, after much deliberation he decided that presenting this chapter from a problem-solving viewpoint and, as much as possible, in an organized tabular form would be of more use to the reader. Consequently, this chapter is short on words and relatively long on tables.

2. TECHNIQUES RELATED TO PROBLEM SOLVING

In the determination as to which statistical analytical methods are to be used for the solution of a given problem it is not so much the statistical problem which determines the technique but rather how the *use* of a given technique is viewed by the problem solver. In other words, it all depends on what the problem solver is trying to accomplish. In Table 1 we show an oversimplified structure of the uses of multivariate analytic methods. Since this table is

Table 1. LIMITATIONS AND CHARACTERISTICS OF STATISTICAL COMPUTER TECHNIQUES

Chapter / Characteristic	3 Subset Selection	4 Stepwise Regression	5 Stepwise Discriminant Analysis	6 Coalitions	7 Factor Analysis Joreskog	8 Factor Analysis Minres
Central memory size						
Low (20K wds)	V	20V	$V \times G = 80$	$V \times G = 50$	$V \times F = 80$	$V \times F = 100$
Medium (60K wds)		100V	$V \times G = 200$	$V \times G = 100$	$V \times F = 200$	$V \times F = 300$
Large ($\geqslant$ 100K wds)		400V	$V \times G = 600$	$V \times G = 400$	$V \times F = 600$	$V \times F = 800$
Word size, bits						
32	P	P	P	P?	P	P
36	P?	P	P	P?	P	P
48–60				P?	P–	
Computational speed, arithmetic						
Floating add $< 1\ \mu$sec	70V					$V \times F = 400$
$\geqslant 1\ \mu$sec	30V	50V	$V \times G = 200$	$V \times G = 150$	$V \times F = 300$	$V \times F = 100$
Logical						
Output graphics	No	Yes	Yes	Yes	No	No
"Cost"	V		$V \times G \times N$	$V \times G \times N$	$V \times F$	$V \times F$
Proportional to						
Solution Uniqueness	No	No	No	No	No	Yes
Optimality	Yes	No	No	Yes	Yes	Yes
Auxiliary storage		Needed	Needed	Needed		

Chapter / Characteristic	9 Unique Rotation	10 Mancova	11 Hierarchical Classification	12 Multidimensional Scaling	13 ISODATA Clustering	14 FFT	15 Forecasting
Central memory size							
Low (20K wds)	V	$N \times V \times F = 200$	40N	$N \times F = 40$	$N \times V = 2{,}000$	100I	100I
Medium (60K wds)		$N \times V \times F = 600$	100N	$N \times F = 100$	$N \times V = 50{,}000$	500I	500I
Large ($\geqslant$ 100K wds)		$N \times V \times F = 1{,}000$	400N	$N \times F = 400$	$N \times V = 200{,}000$	2000I	2000I
Word size, bits							
32	P	P		P	P	P	P
36	P	P?		P?		P?	P?
48–60	P?						
Computational speed, arithmetic							
Floating add $< 1\ \mu$sec	50V	$N \times V \times F = 200$		$N \times F = 100$			
$\geqslant 1\ \mu$sec							
Logical			N		N		
Output graphics	Yes	No	Yes?	Yes?	No	Yes	Yes
"Cost"	$V \times F$	$N \times V \times F$	N	$N \times F$	N	I	I
Proportional to							
Solution Uniqueness	No	?	No	No	No	No	No
Optimality	Yes?	Yes	No	Yes?	Yes?	Yes?	Yes?
Auxiliary storage		Needed	Helpful		Helpful		

F = Factors
G = Groups
I = Intervals
N = Cases
NU = Nonuniqueness of solution
O = Optimality of solution
P = Precision
V = Variables
?= Indicates uncertain importance

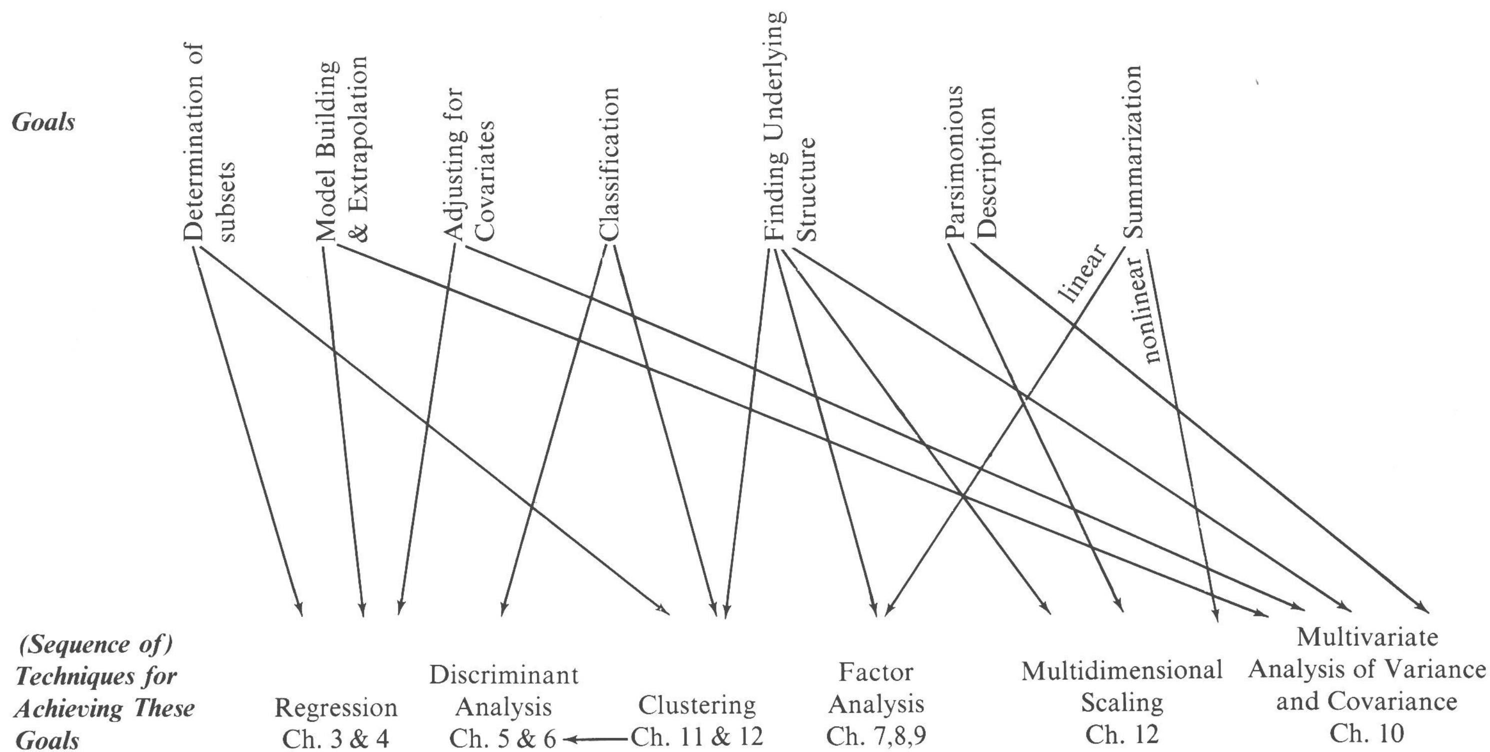


Fig. 1.1. A formulation for data analysis (time series analysis in a separate reference frame).

not a sufficient description of the relationship between methods and goals we have given in Fig. 1.1 a somewhat different viewpoint in which we describe the use of a *sequence* of techniques as contrasted to a single technique for a single task.

A definition of terms used in Fig. 1.1 may be of some aid to the reader:

1. *Model building and extrapolation*: The derivation of an equation or set of equations that explains the variance (variability) in a criterion variable. For example, one observes age at death of a population and wishes to explain the variation in terms of observed parameters such as life habits, attitudes, and genetic background [2].

2. *Adjusting for covariates*: This situation is much more common than is usually appreciated. The aim is to normalize the variables, preferably the dependent variables, in such a way that the effect of disturbing factors that were thought to be controlled but were in fact not controlled, are removed. For example, in the study of a psychotropic drug on which observations were made by a variety of physicians, ranging in specialties from surgeons to psychiatrists, one would find that, if each of these physicians were asked to rate the degree of anxiety of the patient, surgeons would tend to rate the anxiety lower than psychiatrists, with general practitioners and internists falling somewhere in between. In this case, the purpose of covariance adjustment is then to adjust the anxiety judgements of each of the observers in such a way that they become comparable across observers. For this purpose the physician's characteristics, such as his specialty and the size of his practice, are taken into account [3].

3. *Classification* is the assignment of objects to groups. We discuss the topic more thoroughly in the Editorial Preface to the Clustering and Pattern Recognition chapters. However, let it be said now that we call classification using previously established references *learning with a teacher* and using previously unestablished references *learning without a teacher*.

Clustering can be thought of as "learning without a teacher." *Discriminant analysis* in which the groups are preestablished and one wishes to find the separating hypersurfaces between the groups, is "classification with a teacher."

4. *Finding underlying structure*: Given an ensemble of variables or observations, the question is whether the dimensionality of the data is really as large as the number of variables or objects. Is there a subset and, more particularly, a weighted subset of the variables or objects which is able to explain the entire complex within a preset degree of error? The classical approach to this problem has been factor analysis. Multidimensional scaling has come into use in recent years as a nonlinear factor analytic approach. Structure might also imply components of variance and for this reason, of course, multivariate analysis of variance and covariance are often used.

The classical underlying structure problem is found in the use of factor analysis of personality traits, and we would refer the reader to the many works in this area particularly Harman [4] and Chapter 8, and Cattell (Chapter 9) in this volume.

5. *Parsimonious description*: This bears some relation to the finding of underlying structure except that the requirement is to minimize the number of variables used to explain the entire data complex. For example, while all sorts of symptoms may be observed in a biological process, in fact most of the symptoms can be explained by relating them to a very few underlying traits. Here again there is a clear relationship to simple structure.

6. *Summarization*: By summarization we mean something akin to parsimonious description, but extended over multiple populations, for example, the description of personality characteristics in terms of a factor structure or some other multidimensional representation, be it linear or nonlinear.

The last three categories are not very easily separable, yet, in terms of problem solving, they represent clear entities. For this reason, we have chosen to maintain their identity in this chapter.

3. REGRESSION AND DISCRIMINANT ANALYSIS

a. State of the Theoretical and Algorithmic Art

Since the work of Fisher [5] only modest theoretical development has in fact taken place. For regression, most workers have addressed themselves to the problem of selecting the "best" subset of variables. This trend was engendered by the flourishing of numerous, often nonorthogonal variables in many studies. The need, therefore, arose to select those subsets of variables that would best "explain" the variance of the dependent variables in multiple regression, and "best" separate the preidentified groups in multiple discriminant analysis. The methods are generally separable into step-up (or aggregation) and step-down (or backward elimination) procedures. By step-up procedures we mean those in which variables are added to the regression equation one at a time under some criterion, such as: Add that variable which explains the maximum amount of residual variance in the dependent variable, or that which increases the multiple correlation coefficient maximally, and so on. Step-down procedures start with all the variables in the equation and those variables are removed which, for example, will reduce the multiple correlation coefficient the least, either one at a time or several at a time.

There exists presently a lively controversy as to which stepwise methods produce the most robust results. While the most commonly used algorithms aggregate variables [6] (see Chapters 4 and 5) several workers, notably Mantel [7], argue with a great deal of conviction that step-down procedures produce more robust equations. A contender on the scene which might settle the argument is ridge or damped regression [8]. This technique has been used for many years by engineers in nonlinear optimization problems. By the addition of a small damping factor to the diagonal of the correlation or covariance matrix, it is possible to orthogonalize interrelated variables, and by an extension of the technique to treat the overdetermined problem that is, the case in which there are more variables than observations. Ridge regression can then be used as a step-down method and in fact lends further support to Mantel's arguments in favor of step-down procedures.

In discriminant analysis, particularly stepwise discriminant analysis, the original theoretical formulation assumed equal covariance matrices for the separate groups. Recently, Wilf developed the necessary theory and algorithm to permit the treatment of groups with unequal covariance matrices. This method is represented in this book in Chapter 6. This is a distinct advance in the state of the art and in some respects can be equated with the *piecewise linear* discriminant analysis techniques embodied by, for example, ISODATA (Chapter 13), as well as with those which combine the use of linear and piecewise linear techniques in a discrimination problem.

b. Computational Art

ROBUSTNESS

The problem of robustness has also been attacked through the methods for which Tukey coined the term "jackknifing." In multiple regression or discriminant analysis this involves computing estimates of coefficients by removing subsets of the objects or variables with replacement and later calculating an overall estimate of the averaged coefficients. Among the main references in this area, the work of Arvesen and Salsburg [11], Gray and Schucany [12], and the highly readable paper by Mosteller [13] are typical.