ELEMENTS OF STATISTICAL COMPUTING

Numerical computation

Ronald A. Thisted
The University of Chicago

ELEMENTS OF STATISTICAL COMPUTING

Numerical computation

Ronald A. Thisted
The University of Chicago

New York London
CHAPMAN AND HALL

First published in 1988 by
Chapman and Hall
29 West 35th Street, New York NY 10001
Published in the UK by
Chapman and Hall Ltd
11 New Fetter Lane, London EC4P 4EE

Printed in the United States of America

ISBN 0 412 01371 1

Library of Congress Cataloging in Publication Data

Thisted, Ronald A. (Ronald Aaron), 1951–
Elements of statistical computing.
Bibliography: p.
Includes index.
1. Mathematical statistics——Data processing
I. Title
QA276.4.T47 1988 519.5′028′5 86-23227
ISBN 0–412–01371–1

British Library Cataloguing in Publication Data

Thisted, Ronald A.
Elements of statistical computing.
1. Mathematical statistics——Data processing
I. Title
519.5′028′5 QA276.4
ISBN 0–412–01371–1

PREFACE

The purpose of this book is to describe the current state of the art in a number of branches of statistical computing. The very term "statistical computing" is ill-defined, and one aspect of this book is to illustrate the pervasive nature of computation in statistics at all levels, from the most mundane and straightforward applications to the most esoteric and formidable theoretical formulations. What is more, these computational aspects are often at the frontier of knowledge, so that statistical computing now contains a rich variety of research topics in almost every branch of statistical inquiry. A second purpose of this book, then, is to demonstrate the breadth of study and the great room for advances that are possible in areas that simultaneously involve computational and statistical ideas.

Graduate students in Statistics or Computer Science particularly may find this book helpful in suggesting topics worthy of research that would otherwise not be apparent. Researchers in statistical theory and methods may also find this book of value, both for pointing out areas of inquiry that are little known and for the suggestions provided for further reading in each of the areas. In this latter respect, my intent is to provide a source book that can be used as a starting point from which readers can locate results of current research on a wide spectrum of topics in computational statistics. Finally, I hope that this book just makes plain good reading for those who want to stay abreast of advances in this field.

This book makes some assumptions about the background of its reader. The general prerequisite is that attained in the typical first year of graduate study in Statistics, together with a first-year graduate knowledge of computer programming, data structures, and algorithmic analysis. That "typical" first year in Statistics is assumed to provide familiarity both with the basic theoretical ideas and frameworks used in statistics, as well as some experience with fundamental statistical methods and their applications to real problems, that is, to designing real experiments or to analyzing real data. Although this is a *general* level of necessary background, it is not always either necessary or sufficient. Some of the chapters will require deeper understanding of or fuller experience in statistical theory, computer science, or data analysis, while other chapters will require far less.

In particular Chapter 1, which is an overview of the field of statistical computing, is purposely written with a much lower level of background in mind. This chapter assumes the smallest possible exposure to statistical methods and to computation. It can serve as the point of departure for a one-quarter undergraduate course that makes particular use of the experience and background of its students through judicious selection of material. The first chapter also identifies the various threads running through the cloth of the book, and consequently, is the one chapter that serves as introduction to each of the others.

The remaining chapters are meant to stand more or less on their own. Chapter 3, for instance, could serve as the basis for a one semester course in computational aspects of linear models without relying heavily on other chapters. To a lesser extent, I have tried to make the major sections within chapters as independent of one another as possible so as to enhance their value as a reference. Some topics require more space than would be appropriate for a single chapter, and some topics are inherently closely related. I have tried to make these connections natural and obvious.

The chapters of this book could be organized along several different lines. The most obvious choices are discussed at some length in the first chapter, namely, organizing according to branches of statistics (applications, methods, theory, meta-theory), according to branches of computer science (numerical analysis, graphics, software engineering, symbolic computation, artificial intelligence), or according to style of computation (numerical, seminumerical, graphical, symbolic, environmental). Except for Chapter 1, the chapters are ordered more or less according to this last organizing principle, as it seems to me the best way to demonstrate that seemingly disparate topics in statistical computing are really related to one another and lie on a spectrum. In order to present an overview for a statistical audience, however, and to emphasize the pervasiveness of computing in statistics, the first chapter concentrates on illustrating the role of computation in various branches of statistical inquiry.

This is not the first book written on statistical computing. Unlike its predecessors, however, it is not primarily a book of algorithms, nor is it a collection of sources for algorithms. The field has now grown sufficiently that it is now possible to step back somewhat and to attempt an overview which takes account of some of the philosophical issues as well as the purely methodological ones. One fact which makes such an attempt possible is the existence of an excellent work by Kennedy and Gentle (1980) which is a superb compendium of statistical computing methods, emphasizing numerical algorithms. References to Kennedy and Gentle abound here, and with good reason. An earlier book, Chambers (1977), is also notable in this area. Chambers's book is a somewhat broader collection of computational methods relevant to data analysis. At the same time, it is considerably more compact. Although it has been overtaken somewhat by the rapid developments both in computation and in statistics in the last nine years, Chambers's book was the first attempt at a systematic introduction to statistical computing, and it remains valuable.

The emphasis in this book is on computational methods which are of general importance in statistics, either as fundamental building blocks on which other computations are based, or as algorithms which provide insight into the statistical structure of computational problems or the computational structure of statistical problems. There are many computational methods which are highly specialized or highly complex, which do not fall into this category despite their prominence in statistical data analysis (and computing centers' income streams). Examples include computation for the general linear model and for factor analysis. One could easily devote an entire book to the computational issues in either of these areas, indeed, people have done so. What distinguishes these particular techniques is that they are not really building blocks for other statistical computations; rather, they constitute an end in themselves. Many of these computations represent special-purpose elaboration on the more fundamental algorithms and data structures discussed here. Because such excellent programs as SAS and SPSS exist and are generally available, it is unlikely that many statisticians will find the need to write a general analysis of variance program, for example, either for data analysis or for use in their research. Much the same can be said about factor analysis. For this reason, little space is devoted here to the computational details of these procedures. Rather, we have focused on the nature of these computations, and have provided pointers to the literature for the interested reader.

To make this work more suitable for self-study, most sections conclude with a series of exercises based on the material in the section. Answers, hints, or suggestions are provided for many of these exercises. Most of the answers are in outline form only, and require details to be filled in by the reader. It is hoped that the comments are sufficiently detailed to make them useful to the solitary reader while maintaining their utility for classroom use.

Each exercise has been assigned a grade roughly indicating its difficulty. The grading scheme is a Richter-like scale devised by Donald Knuth, in which the rating is approximately logarithmic in the difficulty. Readers who find problems that are either much easier or much more difficult than the difficulty rating suggests are invited to correspond with the author. The following description of the rating system is quoted from Knuth (1968).

Rating Interpretation

00 An extremely easy exercise which can be answered immediately if the material of the text has been understood, and which can almost always be worked "in your head."

10 A simple problem, which makes a person think over the material just read, but which is by no means difficult. It should be possible to do this in one minute at most; pencil and paper may be useful in obtaining the solution.

20 An average problem which tests basic understanding of the text material but which may take about fifteen to twenty minutes to answer completely.

30 A problem of moderate difficulty and/or complexity which may involve over two hours' work to solve satisfactorily.

40 Quite a difficult or lengthy problem which is perhaps suitable for a term project in classroom situations. It is expected that a student will be able to solve the problem in a reasonable amount of time, but the solution is not trivial.

50 A research problem which (to the author's knowledge at the time of writing) has not yet been solved satisfactorily. If the reader has found an answer to this problem, he is urged to write it up for publication; furthermore, the author of this book would appreciate hearing about the solution as soon as possible (provided it is correct)!*

This volume contains the first six chapters of what I once envisioned as a single-volume survey of statistical computing. The outline below of the *Elements of Statistical Computing* contains tentative chapter titles for this lengthier work. It is likely that a second volume containing Chapters 7 through 11 will appear, and occasional references to these chapters crop up in this volume. The field of computing is changing so rapidly, however, that the remainder of the outline may be obsolete by the time the earlier chapters have been finished. For this reason, readers should consider this outline to be the expression of my fond hope that someday these topics will receive an appropriate treatment.

Part I. Introduction and History

1. Introduction to Statistical Computing

Part II. Numerical Computation

2. Basic Numerical Methods

* Donald Knuth, *The Art of Computer Programming*, ©1981, Addison-Wesley, Reading Massachusetts. Reprinted with permission.

3. Numerical Linear Algebra
4. Nonlinear Statistical Methods
5. Numerical Integration and Approximation
6. Smoothing and Density Estimation

Part III. Seminumerical Computation

7. Random Number Generation
8. Monte Carlo Methods
9. Combinatorial and Discrete Methods
10. Experimental Design
11. Discrete Transforms

Part IV. Graphical Computation

12. Standard Graphics
13. Graphical Statistical Methods
14. Graphics Algorithms

Part V. Symbolic computation

15. Fundamentals of Symbolic Computation
16. Algebraic Computation
17. Introduction to Expert Software

Part VI. Computing Environments

18. Examples of Computing Environments
19. Evaluating Statistical Software
20. Data Analysis Environments
21. Other Statistical Computing Environments

Part VII. Theoretical computation

22. Computational Complexity
23. Statistical Algorithms for Parallel Computers
24. Programming Languages

This book was typeset using TEX, the mathematical typesetting system designed and implemented by Donald Knuth (1984). That means that even typogryphical errors are the author's fault. I hope that readers will be kind enough to inform me of any errors they discover, so that they may be expunged from future editions. [Of course, I am in favor of expunging errors, not readers!]

It is a pleasure to acknowledge the support over many years of the National Science Foundation, most recently under Grant DMS-8412233.

Computing equipment used in the preparation of this work was supported in part by NSF Grant MCS-8404941 to the Department of Statistics at the University of Chicago. The manuscript was (nearly) completed while I was on leave at the Department of Statistics at Stanford University. Stanford's hospitality contributed greatly to the completion of this work.

I am indebted to many colleagues whose comments and suggestions upon reading early versions of this material have led to marked improvements and, on occasion, to salvation from committing major blunders. I am particularly grateful in this regard for comments received from Tony Cooper, Sir David Cox, Peter McCullagh, Robert Tibshirani, David Wallace, Sanford Weisberg, and Wing Wong. A number of people responded to a query on the modern history of statistical computation, and I thank them for the recollections and materials that they provided: John Chambers, Sir David Cox, Wilfrid Dixon, James Gentle, John Gower, Peter Huber, and William Kennedy. Any errors of omission or commission are, of course, the author's responsibility.

To my many teachers I owe a great intellectual debt. Although I have received benefit from each of them, I am particularly grateful to four without whom this book arguably would never have come to pass: to Bill Boyd, who first taught me how to write; to Larry Shaw, who taught me how to conduct an experiment; to Isabelle Walker, who introduced me both to statistics and to computation; and to Frederick Sontag, who taught me how to ask questions. Finally, I thank my first and best teachers, my parents.

Ronald A. Thisted
Department of Statistics
The University of Chicago

Chicago, Illinois
September, 1987

CONTENTS

1 INTRODUCTION TO STATISTICAL COMPUTING

It is common today for statistical computing to be considered as a special subdiscipline of statistics. However such a view is far too narrow to capture the range of ideas and methods being developed, and the range of problems awaiting solution. Statistical computing touches on almost every aspect of statistical theory and practice, and at the same time nearly every aspect of computer science comes into play. The purpose of this book is to describe some of the more interesting and promising areas of statistical computation, and to illustrate the breadth that is possible in the area. Statistical computing is truly an area which is on the boundary between disciplines, and the two disciplines themselves are increasingly finding themselves in demand by other areas of science. This fact is really unremarkable, as statistics and computer science provide complementary tools for those exploring other areas of science. What is remarkable, and perhaps not obvious at first sight, is the universality of those tools. Statistics deals with how information accumulates, how information is optimally extracted from data, how data can be collected to maximize information content, and how inferences can be made from data to extend knowledge. Much knowledge involves processing or combining data in various ways, both numerically and symbolically, and computer science deals with how these computations (or manipulations) can optimally be done, measuring the inherent cost of processing information, studying how information or knowledge can usefully be represented, and understanding the limits of what can be computed. Both of these disciplines raise fundamental philosophical issues, which we shall sometimes have occasion to discuss in this book.

These are exciting aspects of both statistics and computer science, not often recognized by the lay public, or even by other scientists. This is partly because statistics and computer science — at least those portions which will be of interest to us — are not so much scientific as they are fundamental to all scientific enterprise. It is perhaps unfortunate that little of this exciting flavor pervades the first course in statistical methods, or the first course in structured programming. The techniques and approaches taught in these courses are fundamental, but there is typically such a volume of material