Statistics and Computing

Springer Science+Business Media, LLC

Statistics and Computing

Gentle: Numerical Linear Algebra for Applications in Statistics.
Gentle: Random Number Generation and Monte Carlo Methods.
Härdle/Klinke/Turlach: XploRe: An Interactive Statistical Computing Environment.
Krause/Olson: The Basics of S and S-PLUS.
Ó Ruanaidh/Fitzgerald: Numerical Bayesian Methods Applied to Signal Processing.
Pannatier: VARIOWIN: Software for Spatial Data Analysis in 2D.
Venables/Ripley: Modern Applied Statistics with S-PLUS, 2nd edition.
Lange: Numerical Analysis for Statisticians.

James E. Gentle

Numerical Linear Algebra for Applications in Statistics

With 21 Illustrations

Springer

James E. Gentle
Institute for Computational
Sciences and Informatics
George Mason University
Fairfax, VA 22030-4444
USA

Series Editors:

Library of Congress Cataloging-in-Publication Data
Gentle, James E., 1943–
Numerical linear algebra for applications in statistics / James E.
Gentle.
p. cm. – (Statistics and computing)
Includes bibliographical references and indexes.
ISBN 978-1-4612-6842-0 ISBN 978-1-4612-0623-1 (eBook)
DOI 10.1007/978-1-4612-0623-1
1. Algebras, Linear. 2. Linear models (Statistics) I. Title.
II. Series.
QA184.G45 1998
512' .5--DC21 98-3959

Printed on acid-free paper.

Production managed by Lesley Poliner; manufacturing supervised by Joe Quatela.
Camera-ready copy prepared from the author's LaTeX files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4612-6842-0

To María

Preface

Linear algebra is one of the most important mathematical and computational tools in the sciences. While linear models and linear transformations are important in their own right, the basic role that linear algebra plays in nonlinear models, in optimization, and in other areas of statistics also makes an understanding of linear methods one of the most fundamental requirements for research in statistics or in application of statistical methods.

This book presents the aspects of numerical linear algebra that are important to statisticians. An understanding of numerical linear algebra requires both some basic understanding of computer arithmetic and some knowledge of the basic properties of vectors and matrices, so parts of the first two chapters are devoted to these fundamentals.

There are a number of books that cover linear algebra and matrix theory, with an emphasis on applications in statistics. In this book I state most of the propositions of linear algebra that are useful in data analysis, but in most cases I do not give the details of the proofs. Most of the proofs are rather straightforward and are available in the general references cited.

There are also a number of books on numerical linear algebra for numerical analysts. This book covers the computational aspects of vectors and matrices with an emphasis on statistical applications.

Books on statistical linear models also often address the computational issues; in this book the computations are central.

Throughout this book, the difference between an expression and a computing method is emphasized. For example, often we may write the solution to the linear system $Ax = b$ as $A^{-1}b$. Although this is the solution (so long as A is square and full rank), solving the linear system does not involve computing A^{-1}. We may write $A^{-1}b$, but we know we can compute the solution without inverting the matrix.

Chapter 1 provides some basic information on how data are stored and manipulated in a computer. Some of this material is rather tedious, but it is important to have a general understanding of computer arithmetic before considering computations for linear algebra. The impatient reader may skip or just skim Chapter 1, but the reader should be aware that the way the computer stores numbers and performs computations has far-reaching consequences. Computer arithmetic differs from ordinary arithmetic in many ways; for exam-

ple, computer arithmetic lacks associativity of addition and multiplication, and series often converge even when they are not supposed to. (On the computer, a straightforward evaluation of $\sum_{x=1}^{\infty} x$ converges!) Much of Chapter 1 is presented in the spirit of Forsythe (1970), "Pitfalls in computation, or why a math book isn't enough."

I emphasize the difference in the abstract number systems called the reals, $\mathbb{R}$, and the integers, $\mathbb{Z}$, from the computer number systems $\mathbb{F}$, the floating-point numbers, and $\mathbb{I}$, the fixed-point numbers. (Appendix A provides definitions for the notation.)

Chapter 1 also covers some of the fundamentals of algorithms, such as iterations, recursion, and convergence. It also discusses software development. Software issues are revisited in Chapter 5.

In Chapter 2, before considering numerical linear algebra, I begin with some basic properties of linear algebra. Except for some occasional asides, this material in the lengthy Section 2.1 is not in the area of *numerical* linear algebra. Knowledge of this material, however, is assumed in many places in the rest of the book. This section also includes topics, such as condition numbers, that would not be found in the usual books on "matrix algebra for statisticians". Section 2.1 can be considered as a mini-reference source on vectors and matrices for statisticians.

In Section 2.2, building on the material from Chapter 1, I discuss how vectors and matrices are represented and manipulated in a computer.

Chapters 3 and 4 cover the basic computations for decomposing matrices, solving linear systems, and extracting eigenvalues and eigenvectors. These are the fundamental operations of numerical linear algebra. The need to solve linear systems arises much more often in statistics than does the need for eigenanalysis, and consequently Chapter 3 is longer and more detailed than Chapter 4.

Chapter 5 provides a brief introduction to software available for computations with linear systems. Some specific systems mentioned include the IMSL Libraries for Fortran and C, Matlab, and S-Plus. All of these systems are easy to use, and the best way to learn them is to begin using them for simple problems.

Throughout the text, the methods particularly useful in statistical computations are emphasized; and in Chapter 6, a few applications in statistics are discussed specifically.

Appendix A collects the notation used in this book. It is generally "standard" notation, but one thing the reader must become accustomed to is the lack of notational distinction between a vector and a scalar.

All vectors are "column" vectors, although I may write them as horizontal lists of their elements. (Whether vectors are "row" vectors or "column" vectors is generally only relevant for how we write expressions involving vector/matrix multiplication or partitions of matrices.)

I write algorithms in various ways, sometimes in a form that looks similar to Fortran or C, and sometimes as a list of numbered steps. I believe all of the descriptions used are straightforward and unambiguous.

One of the most significant developments in recent years, along with the general growth of computing power, has been the growth of data. It is now common to search through massive datasets and compute summary statistics from various items that may indicate relationships that were not previously recognized. The individual items or the relationships among them may not have been of primary interest when the data were originally collected. This process of prowling through the data is sometimes called *data mining* or *knowledge discovery in databases* (KDD). The objective is to discover characteristics of the data that may not be expected based on the existing theory.

Data must be stored; it must be transported; it must be sorted, searched, and otherwise rearranged; and computations must be performed on it. The size of the dataset largely determines whether these actions are feasible. For processing such massive datasets, the order of computations is a key measure of feasibility. Massive datasets make seemingly trivial improvements in algorithms important. The speedup of Strassen's method of an $O(n^3)$ algorithm to an $O(n^{2.81})$ algorithm, for example, becomes relevant for very large datasets. (Strassen's method is described on page 83.)

We now are beginning to encounter datasets of size 10^{10} and larger. We can quickly determine that a process whose computations are $O(n^2)$ cannot be reasonably contemplated for such massive datasets. If computations can be performed at a rate of 10^{12} per second (teraflop), it would take roughly three years to complete the computations. (A rough order of magnitude for quick "year" computations is $\pi \times 10^7$ seconds equals approximately one year.)

Advances in computer hardware continue to expand what is computationally feasible. It is interesting to note, however, that the order of time required for computations are determined by the problem to be solved and the algorithm to be used, not by the capabilities of the hardware. Advances in algorithm design have reduced the order of computations for many standard problems, while advances in hardware have not changed the order of the computations. Hardware advances have changed only the constant in the order of time.

This book has been assembled from lecture notes I have used in various courses in the computational and statistical sciences over the past few years. I believe the topics addressed in this book constitute the most important material for an introductory course in statistical computing, and should be covered in every such course. There are a number of additional topics that could be covered in a course in scientific computing, such as random number generation, optimization, and quadrature and solution of differential equations. Most of these topics require an understanding of computer arithmetic and of numerical linear algebra as covered in this book, so this book could serve as a basic reference for courses on other topics in statistical computing.

This book could also be used as a supplementary reference text for a course in linear regression that emphasizes the computational aspects.

The prerequisites for this text are minimal. Obviously some background in mathematics is necessary. Some background in statistics or data analysis and some level of scientific computer literacy are also required.

I do not use any particular software system in the book; but I do assume the ability to program in either Fortran or C, and the availability of either S-Plus, Matlab, or Maple. For some exercises the required software can be obtained from either `statlib` or `netlib` (see the bibliography).

Some exercises are Monte Carlo studies. I do not discuss Monte Carlo methods in this text, so the reader lacking background in that area may need to consult another reference in order to work those exercises.

I believe examples are very important. When I have taught this material, my lectures have consisted in large part of working through examples. Some of those examples have become exercises in the present text. The exercises should be considered an integral part of the book.

Acknowledgments

Over the years, I have benefited from associations with top-notch statisticians, numerical analysts, and computer scientists. There are far too many to acknowledge individually, but four stand out. My first real mentor — who was a giant in each of these areas — was Hoh Hartley. My education in statistical computing continued with Bill Kennedy, as I began to find my place in the broad field of statistics. During my years of producing software used by people all over the world, my collaborations with Tom Aird helped me better to understand some of the central issues of mathematical software. Finally, during the past several years, my understanding of computational statistics has been honed through my association with Ed Wegman. I thank these four people especially.

I thank my wife María, to whom this book is dedicated, for everything.

I used TEX via LATEX to write the book. I did all of the typing, programming, etc., myself (mostly early in the morning or late at night), so all mistakes are mine.

Material relating to courses I teach in the computational sciences is available over the World Wide Web at the URL,
`http://www.science.gmu.edu/`

Notes on this book, including errata, are available at
`http://www.science.gmu.edu/~jgentle/linbk/`

Notes on a larger book in computational statistics are available at
`http://www.science.gmu.edu/~jgentle/cmpstbk/`

Fairfax County, Virginia

James E. Gentle
June 15, 1998

Contents