Computational Statistics by Geof H. Givens; Jennifer A. Hoeting
Review by: Galin L. Jones
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/27590758
Accessed: 20/10/2014 13:05

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Taylor & Francis, Ltd. and American Statistical Association are collaborating with JSTOR to digitize, preserve and extend access to Journal of the American Statistical Association.
in which the authors discuss common misconceptions regarding Bonferroni’s correction, false discovery rate, and cross-validation. The only fault of the chapter is its length of three pages.

I think the authors have given an excellent treatment of the subject. Comparing the book to its cohort group, it is at a higher level than Draghici (2003) and Simon et al. (2003), but on par with Amarutunga and Cabrera (2004). Unfortunately, it lacks the color graphics found in Draghici (2003), which is almost a necessity for books on this topic. I wish the next edition would provide color graphics. For those considering this book as a course text, be aware that it does not provide exercises at end of each chapter, as do Amarutunga and Cabrera (2004), and to a more limited degree, Draghici (2003). However, textbook exercises in this area are of debatable utility. Although the authors provide their own R package, do not expect much explicit demonstration of R code as in Parmigiani et al. (2003). Nevertheless, software implementation is presented more consistently than Speed (2003) and in much greater detail than Amarutunga and Cabrera (2004).

Overall, I liked this book and would recommend it to any statistician new to microarray data analysis. The book has a unique combination of features that make it a contender among the standard textbooks on microarray data analysis. It is suitable as a text for an introductory course on microarray data analysis. I feel students would enjoy the authors’ straightforward style and their easy-to-read presentation.

J. Wade DAVIS
University of Missouri–Columbia

REFERENCES


Computational Statistics.


Computing is an integral part of modern statistics. Computational intensive methods such as the bootstrap, expectation-maximization (EM) algorithm, and Markov chain Monte Carlo are commonplace in theoretical and applied statistics. As computational speed continues to improve and statisticians attempt to model natural phenomena more accurately, the need for a thorough understanding of sophisticated computational methods can only be expected to increase. This appears to be widely recognized since computing has begun to spread throughout the typical graduate statistics curriculum. Not only are courses on computational statistics flourishing in many departments, but also texts on mathematical statistics frequently cover the use of Mathematica or Maple and introduce important algorithms as well: see, for example, Casella and Berger (2002) and Evans and Rosenthal (2004).

In my own department, a course entitled “Computational Statistical Methods” has been part of the core sequence for (second year) Ph.D. students for several years. The main goal of this course is to provide students with the computational tools useful for conducting research. However, a subtext is that statistical computing is a possible avenue for research in its own right. This means that most of the time is spent on developing good algorithms, methods, and programming habits. For example, since most students will conduct a simulation experiment during their research, an emphasis of the course is appropriate use of Monte Carlo methods. A glance at the Preface and table of contents of Computational Statistics shows that, while there is almost no mention of programming, this book may be appropriate for such a course. The ideal textbook would

- provide exercises with a range of difficulty levels;
- provide a wide range of modern examples but also use many toy examples to illustrate the basic concepts;
- help the reader judge competing methods;
- provide reproducible computational results; and
- provide useful references to additional reading.

The authors state that their goal was to produce a book that will allow one to “develop a broad and thorough working knowledge of modern (emphasis mine) statistical computing and computational statistics.” (For a discussion on the narrow distinction between statistical computing and computational statistics see Gentle 2004.) Perhaps most notably this led the authors to omit any material on pseudo-random number generation and numerical linear algebra. A case could certainly be made that these topics are sufficiently important to be included. However, while I may be inclined to quibble with some of the details, I find that the overall organization and choice of topics is quite good.

The book begins with a review of background material such as Taylor series, likelihood inference, Bayesian inference, statistical limit theory, Markov chains, etc. The remaining 11 chapters form the substantive portion of the book, and these may be roughly divided into four parts: optimization, integration, bootstrap, and smoothing. The optimization section includes rootfinding methods, simulated annealing, genetic algorithms, and a chapter on the EM algorithm. The section concerned with integration covers standard numerical integration techniques such as Gaussian quadrature and then devotes three chapters to Monte Carlo methods including two chapters on Markov chain Monte Carlo. Next is a single chapter on the bootstrap and then three chapters on smoothing and nonparametric density estimation.

I have only a few minor criticisms concerning the structure of the book. First, the scope of material that the authors try to cover in the 17 pages of background material is too broad. I think that the book would be better with Chapter 1 deleted. Second, Monte Carlo methods are introduced in Chapter 6, the bootstrap is introduced in Chapter 9, but Givens and Hoeting make the pedagogically curious decision to discuss application of these methods in the context of EM methods in Chapter 4. However, I do like the fact that in several places the authors included material on using Monte Carlo methods in frequentist settings and, moreover, did not fall into the trap of confusing Markov chain Monte Carlo with “Bayesian computation.” Third, I would like to see the bootstrap chapter greatly expanded.

Generally speaking, Givens and Hoeting write well which results in clearer explanations and fewer typos than normally encountered in a first edition. The one major conceptual error that I found involves their definition of the Gibbs sampler and related examples in Section 7.2. I will not report the details of this error here since there has recently been a second printing in which the authors claim this has been corrected: see Givens and Hoeting (2005). It is obvious that the authors went to great lengths to use many examples based on modern statistical applications throughout the text and to make the data files available on their website. Unfortunately, these examples are sometimes very involved and it is easy to lose track of the point that the authors are trying to make. I would like to see a few more very simple examples used to illustrate the basic points before delving into more substantial examples. The exercises appear well written, thorough, and pedagogically sound. Finally, there are numerous references for the reader interested in obtaining more information about specific topics.

Given the ubiquitous nature of computing in statistics, it should come as no surprise that there are a number of competitors to Computational Statistics. For example, Gentle (2002), Lange (1999), and Monahan (2001) are alternatives that may be suitable for a course on general computational methods. Computational Statistics is a bit more expensive than these alternatives. There are also books on specific topics that may be appropriate. See Fletcher (1987) for material on optimization, Fishman (2000) and Robert and Casella (2004) for Monte Carlo methods, Davison and Hinckley (1999) and Efron and Tibshirani (1993) for the bootstrap, and Eubank (1999) for smoothing and nonparametric regression.

Givens and Hoeting are to be commended for attempting a very ambitious task in writing a book that purports to cover all of the relevant material necessary to “develop a practical understanding of how and why existing methods work... and to provide scientists with the tools [required] to contribute new ideas to the field.” While I think this book falls short of this goal, it does come close and, in the final analysis, could be used effectively in a graduate course on general computational statistics.

Gail L. JONES
University of Minnesota
Developing Statistical Software in Fortran 95.


This book addresses the task of developing statistical applications using modern versions of Fortran. It is primarily directed at the creation of software under the Windows operating system, but some of the material is more generally applicable.

The book begins with an introduction to the new features introduced by the Fortran 90 and Fortran 95 standards. This includes type declarations, control flow, arrays, dynamic allocation, procedures, and modules. The material is covered at an elementary level, and I found it useful to have a Fortran 95 language reference manual on hand to answer the questions which came up as I read.

Next, the book turns to an exposition of "a pseudo-object-oriented style." This includes a discussion of modules, derived types, and pointers. The material in this part of the book is very much aimed at the component object model (COM) client/server model which is introduced later in the book.

The book then addresses the task of implementing computational routines. The subject is addressed in 36 pages which means that the treatment is necessarily abbreviated. The book points out the existence of a number of sources of quality algorithms (e.g., Lapack and the Applied Statistics algorithm series) but also presents a few computational routines, including matrix multiplication, the Cholesky decomposition, and inversion of positive definite matrices. Because the treatment is so abbreviated, anyone reading the book would be well advised to supplement it by dipping into more specialized works on computational statistics.

The second half of the book turns to ways in which Fortran can be used in applications. Three different methods are presented: command-line programs, dynamic libraries, and the Windows COM client/server mechanism. The first two of these methods are relatively straightforward and require only standard system tools, but using Fortran to build COM clients and servers seems to be a relatively complex task. The book gives a clear road map for going about the process and will be a valuable resource for anyone who needs to build COM interfaces using Fortran.

Overall, the book is well written and provides a reasonable introduction to the use of modern versions of Fortran for statistical computation. The real thrust of the book is building COM interfaces using Fortran, and it will no doubt be most useful to anyone who needs to build such interfaces.

Ross Ihaka
The University of Auckland

REFERENCES


An R and S-Plus® Companion to Multivariate Analysis.


This book is designed to complement another more theoretical textbook in multivariate analysis for undergraduate or graduate students in applied statistics courses. In particular, the content of Everitt’s text is compatible with material presented by Efron and Dunn (2001), Johnson and Wichern (2003), and Rencher (2002). It also would serve as an excellent additional reference for the applied statistician confronted with multivariate data.

The book begins with an introduction to multivariate data and suitable graphical displays. This is followed by principal components and factor analysis, multidimensional scaling and correspondence analysis, cluster analysis, MANOVA and discriminant analysis, multiple regression and canonical correlation, and finally a chapter on repeated measures data. A rudimentary acquaintance with matrix algebra is presumed, and there is no self-contained introduction or review. Everitt also expects the reader to have some familiarity with S-PLUS or R, although a brief overview of basic R/S commands is given in the Appendix.

A companion website from which the reader can download the datasets and R/S code used in the examples is provided by the author so that the reader can easily replicate the analyses presented in the text. Throughout the textbook, Everitt is careful to identify any differences in the code necessary for R versus S-PLUS implementations and does not cater to either software more or less favorably.

This text is much more than just an R/S programming guide. Brian Everitt’s expertise in multivariate data analysis shines through brilliantly. Each chapter roughly has the following outline: introduction to the statistical method, motivated by a scientific research problem; bulleted display summarizing the theory and/or procedure for the analysis under consideration (often quite detailed); at least one example of an application using R and S-PLUS; summary; and exercises. Though the book does not delve much into the theory underlying the statistical methods presented, it is readable on its own, and provides a self-contained reference for multivariate data analysis. Everitt’s detailed treatment of the statistical methodology itself does, however, mean that less space is devoted to discussing the R/S functions and their various options and parameters. The reader will undoubtedly need to supplement this text with S-PLUS or R documentation for the multivariate data analysis functions themselves.

There are a few differences between Everitt’s coverage of material and what is included in other introductory books. Chapter 2, for instance, “Looking at Multivariate Data,” does not discuss faces or star plots, although these are available in the standard S-PLUS library of functions. Instead, Everitt provides a detailed account of variation, and extensions of the scatterplot, bivariate density estimation, an introduction to trellis and conditional graphs, and three-dimensional graphs. The chi-plot is presented as a means of graphically assessing the independence of two quantitative variables (Fisher and Switzer 2001), and the bivariate boxplot is also introduced (Goldberg and Iglewicz 1992). Functions to create these two types of graphs are provided at Everitt’s companion website, as they are not standard features of S-PLUS and R. It should also be noted that the chapter “Multiple Regression and Canonical Correlation” (Chap. 8) does not include a discussion of multivariate multiple regression but rather reviews the multiple correlation coefficient and univariate multiple regression to then transition into canonical correlation.

The treatment of repeated measures data analysis in Chapter 9 is limited to an approach via linear mixed effects models. In particular, the random slope, and random intercept, and slope models are introduced. Other approaches to repeated measures analysis, profile analysis, and growth curves are not mentioned. Therefore, the tests of hypotheses related to these latter approaches (e.g., tests regarding parallel, flat, or coincident profiles), which are presented in other multivariate texts are not discussed here and hence neither are their R/S implementations. Everitt concludes this chapter with a nice section on distinguishing different types of missing data and comments on the consequences of missing data in a longitudinal analysis.

The author may wish to consider adding an errata page to his companion website, as several tables presented in the latter chapters of the book contain typographical errors. None of these seriously detract from the readability of the text but could cause a little confusion for beginning students.