James E. Gentle: Computational statistics

Christian Robert

Computational Statistics, James E. Gentle, April 2009, Statistics and Computing Series, Springer-Verlag, New York, xxi+727 pages, ISBN 978-0-387-98143-7, DOI 10.1007/978-0-387-98144-4, $ 99.00

This book has a very large scope in that, beyond its title, it covers the dual fields of computational statistics and of statistical computing. If only for this reason it thus constitutes a reference text for any statistics course including some degree of computing. The first field, computational statistics, focus on statistical methods that rely heavily on computational tools ("Computational inference, together with exact inference and asymptotic inference, is an important component of statistical methods"), like the Bayesian analysis of hierarchical models (Gelman et al., 2001), while the second field covers computational or numerical methods that are used by computational statistics, like the EM algorithm (Dempster et al., 1977). The latter field being more easily chartered and more operational than conceptual, it may appear to the reader that, paradoxically, *Computational Statistics* focus more on statistical computing than on computational statistics.

Computational Statistics is structured as a four part treatise:

- **Part I: Mathematical and statistical preliminaries**
- **Part II: Statistical Computing** (Computer storage and arithmetic.- Algorithms and programming.- Approximation of functions and numerical quadrature.- Numerical linear algebra.- Solution of nonlinear equations and optimization.- Generation of random numbers.)
- **Part III: Methods of Computational Statistics** (Graphical methods in computational statistics.- Tools for identification of structure in data.- Estimation of functions.- Monte Carlo methods for statistical inference.- Data randomization, partitioning, and augmentation.- Bootstrap methods.)
- **Part IV: Exploring Data Density and Relationship** (Estimation of probability density functions using parametric models.- Nonparametric estimation of probability density functions.- Statistical learning and data mining.- Statistical models of dependencies.)

some entries (like bootstrap or Monte Carlo) being found in more than one part.

The first part of *Computational Statistics* is, as indicated by its title, a single preliminary chapter containing what Gentle considers as the essentials of mathematics, probability, and statistics for the understanding of the

book. A reader unfamiliar with too many topics within this chapter should first consider improving his or her background in the corresponding area! This is a rather large chapter, made of 82 pages, and it offers some usefulness for all readers, not only to signal deficiencies in their background, as noted above, but also to anchor the subsequent developments into the proper formalism. Given this purpose, the few exercises in this chapter are necessarily incomplete but they can be seen as useful reminders. (However, some of those exercises involve simulation tools introduced later in the book and thus should be set aside for future processing). Gentle insists in his preface on the urgency of reading this chapter, even for seasoned statisticians, for fear of missing the "subtle points". (For instance, he stresses the difference between the "form of a mathematical expression and the way the expression should be evaluated in actual practice", illustrating this point with the standard example that solving $Ax = b$ in x does not require computing A^{-1}.)

As shown by the titles of the chapters therein, the second part of *Computational Statistics* is truly about the theory of computation, meaning using computers for numerical approximations, with interesting sections about the representation of numbers in computers, about approximation errors, and of course about random number generators. While there exist complete books about everyone of those topics (see, e.g., Fishman, 1996), I appreciate the need for enforcing upon our students' minds those hardware subtleties, especially since they often seem completely unaware of them, despite their advanced computer skills. (I like very much the repeated aphorism "Programming is the best way to learn programming".) This second part of *Computational Statistics* may thus appear as a self-contained textbook of 250 pages on numerical methods (incl. function approximations by basis functions, resolution of linear and non-linear systems) and on random generators, i.e. it covers the same ground as the set of Gentle's earlier books, *Random Number Generation and Monte Carlo Methods* (2004) and *Numerical Linear Algebra for Applications in Statistics* (1998), while *Elements of Computational Statistics* (2002) looks very much like a shorter entry on the topics treated in Parts III and IV of *Computational Statistics*. This second part can certainly and seamlessly sustain a whole semester undergraduate course while advanced graduate students can be expected to gain from a self-study of those topics. In my opinion, it is a very coherent part of the book and it constitutes a must-read for all students and researchers engaging into any kind of serious statistical programming. Obviously, some notions are introduced a bit superficially, given the scope of this section (as, for instance, MCMC techniques, in less than six pages), but I came to realise this

is the very point of the entire book, namely to provide an entry into "all" necessary topics, along with links to the relevant literature (if often entries in Gentle et al., 2004) for deeper enlightment. As such, I deplore that the ultimate issue of Monte Carlo experiments, whose construction often is a hardship for students, is postponed till the 100 page long appendix. (In my experience, students do not read appendices!)

The third and fourth parts of *Computational Statistics* cover the major methods of computational statistics, including Monte Carlo methods (I appreciated very much the quote "Monte Carlo methods differ from other methods of numerical analysis in yielding an estimate rather than an approximation"), bootstrap, and randomization, but also non-parametric methods like probability density estimation. Because of this wide scope, Part III is to be understood as an introduction to the tools of the trade. As a result, the depth with which the different topics are processed is variable, and most chapters are rather short even though they try to be all-encompassing. Part IV is a description of machine learning techniques, that could well function as a preliminary to Hastie et al. (2010).

In my opinion, some relevant sections of Part III would have fitted better in Part IV, where they belong. For instance, Chapter 10 on the estimation of functions covers the evaluation of estimators of functions, but postpones the actual construction of those estimators till Chapter 15. Jackknife is introduced on its own in Chapter 12, while the bootstrap is covered in Chapter 13. (I note that bootstrap for non-iid data is dismissed rather quickly, given the current research in the area.) The first sections of Chapter 16 are a valuable entry on clustering and data-analysis tools like PCA, while the final section of this chapter on high dimensions feels slightly out of context. (I also bemoan that the *curse of dimensionality* is mentioned too late in the book.) Chapter 17 is about dependent data and discuss regression, classification and projection pursuit, but does not cover the existing literature on graphical models and their use in the determination of dependence structures.

In conclusion, *Computational Statistics* is a very diverse book that can be used as a textbook at several levels of a statistics curriculum, as well as a reference for researchers (even if mostly as an encyclopedic entry towards further and deeper references). The book is well-written, in a lively and personal style. There is no requirement for a specific programming language, algorithms being described in a pseudo-code manner. The R language is introduced in a somewhat dismissive way (R "most serious flaw is usually lack of robustness" since "some [packages] are not of high-quality") and the software BUGS (Lunn et al., 2000) is not mentioned at all. (Some exercises

are of the kind "Design and write either a C or a Fortran subroutine", when they should simply ask for a program in any language.) The appendices of *Computational Statistics* also contain the solutions to some exercises, even though the level of detail is highly variable, ranging from one word ("1") to a whole page (see, e.g., Exercise 11.4). The twenty-some page list of references is preceded by a few pages on currently available journals and webpages. Despite the reservations raised above about some parts of *Computational Statistics* that would benefit from a deeper coverage, I think this book is a reference book that should appear in the shortlist of any computational statistics/statistical computing graduate course as well as on the shelves of any researcher supporting his or her statistical practice with a significant dose of computing backup.

References

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Society Series B*, 39:1–38.

Fishman, G. (1996). *Monte Carlo*. Springer-Verlag, New York.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2001). *Bayesian Data Analysis*. Chapman and Hall, New York, New York, second edition.

Gentle, J., Härdle, W., and Mori, Y. (2004). *Handbook of Computational Statistics—Concepts and Methods*. Springer-Verlag, Berlin.

Gentle, J. E. (1998). *Numerical Linear Algebra for Applications in Statistics*. Springer-Verlag, New York.

Gentle, J. E. (2002). *Elements of Computational Statistics*. Springer–Verlag, New York.

Gentle, J. E. (2004). *Random Number Generation and Monte Carlo Methods*. Springer-Verlag, New York, 2nd edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2010). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, New York, second edition.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comput.*, 10:325–337.