

Title: Computational Statistics Handbook with MATLAB
Authors: Wendy L. Martinez and Angel R. Martinez
Chapman & Hall/CRC, Boca Raton, 2002. pp xvii + 591

Reviewed by: Byron J.T. Morgan
b.j.t.morgan@kent.ac.uk

The computer revolution has greatly influenced how we do and teach statistics. Thus one now often encounters university lecture courses on Computational Statistics. Such courses typically cover a wide range of techniques and are often linked to appropriate statistical packages where students can carry out illustrative analyses themselves. This book may be used for such a course, making use of the technical computing environment of MATLAB.

MATLAB is a popular and powerful computing environment, which is also available in a Student edition. It has an attractive language, and possesses especially good graphics. The web-site <http://www.mathworks.com/support/books> lists over 600 books for use with MATLAB and related packages. The basic functionality of MATLAB is extended through a library of separate Toolboxes, such as the Optimization Toolbox, the Signal Processing Toolbox, and so forth. Appendix E of this book describes the list of functions available in the Statistics Toolbox. Although the Statistics Toolbox is now quite extensive, there are a number of gaps, and many of these are filled by the programs of the book, which are listed in the book Appendix F, and may be easily downloaded, as a Computational Statistics Toolbox, from the site: <http://www.infinityassociates.com>. There are also overlaps with the Statistics Toolbox. Also available are the many data sets that are used to illustrate techniques in the book, documented in Appendix G; several of these are taken from Hand et al (1994).

The book aims to promote the use of MATLAB by statisticians, and to extend the computational statistics functionality of the MATLAB environment. The target audience is general scientists, such as engineers and psychologists, and senior undergraduates and postgraduate students in statistics and engineering. In terms of prior knowledge, in order to follow the material of the book, it helps to have some experience of probability and statistics, as well as linear algebra (because MATLAB is array-based). An attractive feature of the book is the large number of exercises at the end of each Chapter, which are a valuable source of ideas for lecturers in general courses on Computational Statistics. Some of the exercises are of the general type: "Write a MATLAB program to...", but others involve innovative and imaginative use of computing. Appendix A of the book provides an Introduction to MATLAB, highlighting, for example, the useful element-wise operators for arrays. The book uses MATLAB Version 6, and version 3 of the Statistics Toolbox. Of course, packages change regularly, and Version 4 of the Statistics Toolbox is described at: <http://www.mathworks.com/products/statistics>.

The book is dedicated to Edward Wegman, and in Chapter 1, Computational Statistics is defined, from Wegman (1988), as a collection of techniques that have a strong focus on the exploitation of computing in the creation of new statistical methodology. It is explained that the emphasis in the book is on algorithmic developments, and demonstration of how techniques work, omitting much of the background theory. As will be discussed later, this sometimes means that some topics are covered superficially. However, each chapter ends with a Further Reading section, which provides up-to-date comprehensive references for those who want to check up on theory, and also to alternative computational resources. Overall, this is very well done, and the balance between coverage and detail is quite good. For instance, simple algorithms are described in Chapter 4 for simulating gamma and beta random variables with integer indices, but the exercises consider more general procedures. Chapters 2 and 3 of the book cover basic probability and sampling concepts, with short MATLAB programs for illustration as appropriate. This material is pretty basic, and should be easily understood. Inevitably, coverage is often quite brief, for instance Maximum likelihood estimation is covered in just a few pages, and with no discussion of iterative methods for solving systems of partial-derivative equations. Chapter 4 is about simulating random variables, in recognition of the importance of this topic to Computational Statistics. The emphasis here is on inversion and rejection methods, so that there is, for instance, no mention of the ratio method. Chapter 5 is on exploratory data analysis, and covers a wide range of useful procedures, ranging from boxplots and scatterplots to Projection Pursuit and the Grand Tour. Chapter 6 covers Monte Carlo methods, such as the Bootstrap and Monte Carlo hypothesis testing, and an interesting focus of Chapter 7 is on Data partitioning, including cross-validation and the jackknife. It is here, under the section on cross-validation, that one is introduced to linear regression. Chapter 8 covers probability density estimation, with standard material on histograms, kernel density estimation and finite mixtures (which is where to find discussion of the EM algorithm). Under the title, Statistical Pattern Recognition, Chapter 9 covers material such as Receiver Operating Curves, classification trees and clustering. The example used to illustrate dendrograms has just 5 objects, and it

is a pity that a more interesting illustration was not used here. Chapter 10 covers Nonparametric regression, including Loess curve construction and the Nadarya-Watson estimator, but omitting spline-based procedures. The subject of Chapter 11 is the basics of Markov chain Monte Carlo, including the Gelman-Rubin method for monitoring convergence, and the Raftery-Lewis method, which is available in the Econometrics Toolbox. The last chapter is on Spatial statistics, as “an area of data analysis where the methods of computational statistics can be applied”, and topics covered include visualising and simulating spatial point processes and estimating spatial dependence. There is no mention of kriging.

The writing is clear, and the book is a pleasure to read. It also works well as a source of reference. The MATLAB programs are well-commented, which makes them easy to follow and understand. Simple functions are included in the code through use of the *inline* facility, which also aids comprehension. The programs are not intended to be efficient, as there is an emphasis on transparency and understanding.

Three examples of where the material suffers from brevity are as follows:

- In the derivation of the Normal Reference Kernel, there is a confusion between the form of the kernel and the assumed form of the probability density function, which makes it hard for the reader to understand how the stated standard result is achieved.
- The Metropolis-Hastings sampler is illustrated for generating a random sample from a Cauchy distribution. Here the acceptance function is taken to be a normal probability density function, so that the method is in fact the Metropolis sampler (which is formally discussed in the following section of the book), and the associated code is too complex as a result.
- When hierarchical cluster-analysis is discussed, there is a common confusion between method and algorithm: “*There are two types of hierarchical clustering methods: agglomerative and divisive*”. The single-link method, for instance, can have

both agglomerative and divisive algorithms. Single-link cluster-analysis receives the usual criticism, that the resulting clusters may suffer from chaining. However other hierarchical methods can be non-unique, and may produce inversions (see eg., Morgan and Ray, 1995), and this is not mentioned.

In terms of coverage, it is surprising that there is no discussion of material on generalised linear modelling, which should be a staple component of courses on Computational Statistics. It is interesting to compare coverage of topics with that of the book by Venables and Ripley (1999), which does cover generalised linear models, as well as GAMs, survival analysis and time-series, all of which are missing here. Of course, in a book of this length, it is inevitable that the authors have to be selective in what they decide to cover.

I give a lecture course for senior undergraduates/MSc students on Computational Statistics in which techniques are also illustrated by MATLAB programs (see Morgan, 2000). The students generally seem to like the approach, and find the computer environment to be friendly and useful. It should come as no surprise therefore that I found this a very enjoyable book to read, and that it has been added to the reading list for the lecture course in question. Overall, this book succeeds in its two major aims, and should be widely used as a source of reference, as well as for use on relevant lecture courses.

References:

- [1] Hand, D, Daly, F, Lunn, A D, McConway, K J and Ostrowski, E (1994) *A Handbook of Small Data Sets*, London: Chapman & Hall
- [2] Morgan, B J T (2000) *Applied Stochastic Modelling*, London: Arnold
- [3] Morgan, B J T and Ray, A P G (1996) Non-uniqueness and inversions in cluster analysis. *Applied Statistics*, 44, 117-134
- [4] Venables, W N and Ripley, B D (1999) *Modern applied statistics with S-Plus (3rd edition)* New York: Springer
- [5] Wegman, E (1988) Computational statistics: a new agenda for statistical theory and practice. *J. Washington Ac. Sci.*, 78, 310-322