



Journal of Statistical Software

April 2005, Volume 14, Book Review 2.

<http://www.jstatsoft.org/>

Reviewer: Duncan Murdoch
University of Western Ontario

Computational Statistics

Geof H. Givens and Jennifer A. Hoeting
John Wiley & Sons, Hoboken, New Jersey, 2005.
ISBN 0-471-46124-5. xix+418 pp. \$89.95 (P).
<http://www.stat.colostate.edu/computationalstatistics/>

This is a graduate-level text aimed at developing a “broad and thorough working knowledge of modern statistical computing and computational statistics”, to quote the preface.

The book consists of 12 chapters, in 5 groups: review, optimization, integration (including Monte Carlo), bootstrapping, and smoothing:

1. Review
2. Optimization and Solving Nonlinear Equations
3. Combinatorial Optimization
4. EM Optimization Methods
5. Numerical Integration
6. Simulation and Monte Carlo Integration
7. Markov Chain Monte Carlo
8. Advanced Topics in MCMC
9. Bootstrapping
10. Nonparametric Density Estimation
11. Bivariate Smoothing
12. Multivariate Smoothing

The coverage of these topics is very complete, including advice and citations of authors who offer dissenting advice. The writing style is very clean and readable. Nontrivial exercises are given at the end of each chapter. There are few typos, with an errata sheet available.

The authors have intentionally omitted some topics when they thought that existing software solved the problem so well that students need not know the details: pseudo-random number generation for standard distributions (though rejection sampling is described in detail) and numerical linear algebra are the two main examples. They cover a very wide range of other modern topics, with 581 references. The references are quite current up to 2002, and later in a few cases.

To give a flavour of the text, here are detailed descriptions of the contents.

Chapter 1 contains a concise review of mathematical limit theory, statistics including likelihood-based and Bayesian inference, statistical limit theory, and Markov chains.

Chapter 2 is about solving smooth nonlinear equations, mainly by solving $f'(x) = 0$ in a univariate or multivariate context. Newton's method, Fisher scoring, the secant method, and fixed point iteration are developed for univariate methods. The order of convergence is defined and worked out for these algorithms. For multivariate problems, a variety of methods are described, including Newton-like methods, steepest ascent, discrete Newton, quasi-Newton, Gauss-Newton and Gauss-Seidel. My only complaint about this chapter was that methods which are more robust in the face of non-smooth objective functions (the golden-section search, the Nelder-Mead simplex algorithm) are given short shrift.

Chapter 3 "Combinatorial Optimization" discusses optimization problems on discrete spaces. After developing the notation, the authors discuss the concept of the order of difficulty of a problem: the classes P, NP, NP-complete and NP-hard are defined. The general topic is illustrated with an example on genetic mapping, and the need for heuristics is motivated. Local search algorithms are defined, and another example (on model selection for baseball salaries) is used to illustrate. Tabu algorithms are discussed at length (and applied to the baseball data). Simulated annealing and genetic algorithms follow, both with extensive discussion and application to the baseball data.

Chapter 4 is about the EM algorithm, which is well-illustrated by worked examples; variations are given.

Chapter 5 on numerical integration goes into perhaps a little too much detail on one-dimensional integration, covering Newton-Côtes quadrature, Romberg integration, Gaussian quadrature. It finishes with a brief mention of adaptive quadrature and symbolic integration. My feeling is that most readers do not need this level of detail on one-dimensional integration: they will either be fine with a canned package to handle it, or will need to consult a more advanced text.

Chapter 6 briefly discusses Monte Carlo integration in general, then gives a table of methods for generating random variables from standard parametric families. Rejection sampling is discussed in some detail, including variations such as adaptive methods. (Here the authors make a small slip: they say that when extending adaptive rejection sampling to densities that are not log-concave, MCMC is necessary. This is not true; it is just what Gilks *et al.* (1995) chose to do. Other adaptive algorithms are possible.) Importance sampling is developed, with perhaps a little too much emphasis on using sampling importance resampling (SIR) to generate approximate i.i.d. samples from the target distribution. The point of the chapter is integration; the weighted importance sample is better for that. Bridge and path sampling are briefly discussed, as are antithetic sampling, control variates, and Rao-Blackwellization. This is an excellent chapter.

Chapter 7 "Markov Chain Monte Carlo" discusses a variety of Metropolis-Hastings algorithms including the independence sampler, random walk chains, hit-and-run algorithms, Langevin chains, and multiple-try Metropolis-Hastings. It then discusses variations on the Gibbs sampler. This is perhaps the only disappointing section of the text. The Gibbs sampler is first implemented with component i updated to time $t + 1$ conditional on the values of the other components at time t , rather than the standard implementation where component i is updated conditional on the most recent values of all other components. This implementation

does not work: the limiting distribution can be different than when it is done correctly. The authors do go on to recommend the standard implementation (which they give later under the name “immediate updating”), but readers might get the impression that both algorithms are valid. I would much prefer the first version to be presented with a clear “this does not work” warning. The Gibbs sampler section again disappoints with the illustrative example. A hierarchical model is developed and then simulated using Gibbs. But direct simulation from the model would produce an i.i.d. simulation, so why go to the trouble to use Gibbs? Conditioning on some of the component values in the model (e.g. the observed data in a Bayesian hierarchical model) is enough to make direct simulation difficult; this is the situation where Gibbs is useful.

After the different types of chains are described, practical implementation issues are discussed, including the choice of proposal in Metropolis-Hastings, identifying burn-in time, and making use of the results. A detailed example based on capture-recapture data is presented.

Chapter 8 covers advanced topics in MCMC: auxiliary variable methods including the slice sampler; Green’s reversible jump MCMC, and perfect sampling (CFTP). (There’s another slip in the introduction to the reversible jump algorithm: the authors claim that a chain cannot be reversible if $X^{(t)}$ and $X^{(t+1)}$ have different dimensions. But of course, the whole point of RJMCMC is to construct a chain which is reversible and does dimension-changing jumps.) A detailed example illustrating MCMC for Markov random fields is given. This is another excellent chapter.

Chapter 9 discusses bootstrapping. It first introduces the ideas, then spends some time on bootstrap inference, concentrating on the percentile method for confidence intervals at first. I would rather have seen the “basic bootstrap” confidence interval first; it has a more natural derivation without requiring a symmetry assumption. Unfortunately, the basic bootstrap is not discussed at all, though other variations are: the accelerated bias-corrected method, the bootstrap t , empirical variance stabilization, prepivoting, the balanced and antithetic bootstraps. A summary of approximation results for bootstrap convergence is given, but not derived. The chapter finishes with permutation tests.

The last three chapters in the book deal with smoothing. Chapter 10 is about density estimation, mainly concentrating on kernel-based methods. A careful discussion of bandwidth and kernel selection is included. Logspline methods are also discussed, followed by a number of multivariate methods including a long example using projection pursuit.

Chapter 11 covers scatterplot smoothing and (briefly) general bivariate curve fitting. The ideas of equivalent kernels are developed for linear smoothers; loess and supersmoother are given as examples of nonlinear smoothers. Confidence bands are developed.

The final chapter surveys a number of topics in multivariate smoothing: additive models, generalized additive models, projection pursuit regression, neural networks, alternating conditional expectation, etc. Tree-based methods are covered in a fair amount of detail. The chapter finishes with a short description of principal curves in multivariate data.

This book includes a solid theoretical background at the introductory graduate level, practical advice, application to real datasets, and very few errors. It covers a large selection of topics very well. The book is not perfect, but besides the problems with Gibbs sampling described above, my complaints with it are very minor. This is an excellent first edition of a text that I hope to use the next time I teach a statistical computing course.

Reviewer:

Duncan Murdoch

Department of Statistical and Actuarial Sciences

University of Western Ontario

London, Ontario, Canada N6A 5B7

E-mail: murdoch@stats.uwo.ca

URL: <http://www.stats.uwo.ca/faculty/murdoch/>