Journal of Statistical Software

August 2015, Volume 66, Book Review 3. *http://www.jstatsoft.org/*

Reviewer: Abdolvahab Khademi
University of Massachusetts

Statistical Modeling and Computation

Dirk P. Kroese and Joshua C. C. Chan
Springer-Verlag, New York, 2014.
ISBN 978-1-4614-8774-6. xx + 400 pp. USD 79.99 (H).
http://www.springer.com/book/9781461487746

The field of statistics has transformed in several ways in the past decades. Two changes are notable: (1) a synergy between the statistics field proper and fields that not only use but also contribute to the field, including agriculture, psychology, biostatistics, econometrics, and psychometrics, among others; and (2) the exploitation of computing power both in enhancing statistical analysis and modeling and also design of new algorithms in implementing statistical methods. Nonetheless, the core of the field, which builds on probability theory, theoretical statistics and mathematical statistics, remains as fundamental, for which existing standard texts provide solid foundations. What is left desired in these texts, however, is the incorporation of the two foregoing changes, namely a more practice-oriented approach and a more computational treatment of the concepts, the mathematics and the models. *Statistical Modeling and Computation* is a text that aims to fulfill this promise.

As the title of the book suggests, there is a computation aspect to the text. The authors have chosen MATLAB as the primary computing language throughout the book, with corresponding R code on the book companion website (which also includes MATLAB code and a small errata document). The choice of MATLAB language can be presumed to be a hint at the mathematical and numerical underpinning of the book.

The book is written in eleven chapters, most with extensive mathematical, applied and computational exercises and problems with selected solutions at the end of the book. The eleven chapters are appropriately grouped into three main parts: (I) fundamentals of probability, (II) statistical modeling and classical and Bayesian inference, and (III) advanced models and inference.

Part I of the book, *Fundamentals of Probability*, includes three chapters. In Chapter 1 (*Probability Models*), the authors present the fundamentals in probability and mathematical statistics, such as random experiments, events, sample space, and conditional distribution. This is taken further in Chapter 2 (*Random Variables and Probability Distributions*) which presents common discrete and continuous distributions with their mathematical forms, properties, proofs and graphs. The authors also demonstrate small simulations in generating random variables based on some of those distributions. Chapter 3 (*Joint Distributions*) smoothly extends the concepts in the previous two chapters to multivariate cases, hence presenting concepts such as discrete joint distributions, continuous joint distributions and mixed joint distributions in addition to the concepts of covariance, linear and general transformations (such as Box-Muller method), multivariate Gaussian distribution (and a small simulation code for that) and central limit theorem. While Part I reinforces probability concepts, Part II introduces statistical modeling.

Part II, *Statistical Modeling and Classical and Bayesian Inference*, constitutes the core of the text with five chapters. Chapter 4 (*Common Statistical Models*) begins this part with a very insightful yet simple flowchart that depicts the entire field of statistical modeling, conceived as building a working and precise model that can capture the reality through collected data. This chapter briefly discusses linear models manifested as (multiple) regression and ANOVA and then recasts both regression and ANOVA as normal linear models. Although this chapter is rather brief, it compensates for that by providing excellent model-thinking problems using different scenarios. In Chapter 5 (*Statistical Inference*) the authors start out with distinguishing between classical and Bayesian statistics and proceed with the former school of thought, leaving the other to be treated in a separate chapter. This chapter attempts to present statistical inference as mainly estimation of unknown parameters. Methods of moments, least-squares and maximum likelihood estimation are briefly presented as major point estimation methods, the latter of which is considered more powerful and which takes a dedicated separate chapter. In addition to point estimation procedures, the authors present confidence interval estimation methods for discrete and continuous distributions for both one- and two-sample data. MATLAB code is provided for some examples for both point and interval estimations. Next, the authors introduce hypothesis testing in classical statistics with the apt caveat that modern statistical analyses, especially computation-intensive ones, tend to use confidence intervals or the Bayes factor instead. This is a very good point regarding modern statistical analysis, especially in statistical learning applications. Hypothesis testing for linear models is presented as model selection, followed by cross-validation methods of k-fold and leave-one-out procedures with brief introduction of training and test data/error and sample MATLAB code to implement LOO-CV.

Chapter 6, (*Likelihood*), presents the likelihood approach to parameter estimation starting out with examples based on binomial and normal distributions. The authors present the score function, which is used in their treatment of Fisher information and Cramer-Rao inequality. Maximum likelihood estimation (MLE) is then presented as a parameter estimation method and exemplified by example data from binomial, iid normal, normal linear and exponential families. Score sets are also shown for constructing confidence interval sets for MLE calculated parameters. Some mathematical properties of the ML estimator are given (such as invariance, consistency and asymptotic distribution) before introducing the application of likelihood in test statistics. As for numerical methods to find the maximum likelihood estimator, the authors discuss the Newton-Raphson iterative method and the three-step expectation-maximization (EM) method, which is more widely used in latent variable modeling and mixture models. Both methods are introduced with examples and MATLAB code, although the EM section receives a more in-depth treatment and elaboration.

Simulation provides researchers with tools to study different properties of statistical models and parameters under different conditions. This important topic is taken up in Chapter 7 (*Monte Carlo Sampling*) with a focus on bootstrap methods and Markov chain Monte Carlo

(MCMC) methods. The authors show how to use empirical CDF and density estimation to generate iid samples. Bootstrapping is demonstrated in estimating regression weights and ratio estimators. Metropolis-Hastings and Gibbs sampling methods are introduced in detail with appropriate MATLAB code, examples and related graphs. Because this chapter deals with simulation, naturally it includes more computation exercises than other chapters in the book, presenting a flavor of statistical computing to the audience. For those who need practice learning simulation, the exercises in this chapter provide an excellent opportunity.

Bayesian statistics is presented in detail in Chapter 8 (*Bayesian Inference*), distinguished from frequentist approach through definitions and use of different notations. Some of the examples from classical chapters are rewritten using the Bayesian approach, providing the reader a tangible understanding of these different approaches. Normal, linear, and multinomial models are presented and mathematically elaborated as the most common Bayesian models. Bayesian networks are also introduced together with a practical example in classification contexts. The Bayes factor is described in detail for Bayesian model selection with illustrative examples and MATLAB code. This chapter ends Part II of the book, preparing the audience for the more applied Part III.

Part III, *Advanced Models and Inference*, encompasses three chapters on generalized linear models (GLM), dependent data models, and state space models. Chapter 9 (*Generalized Linear Models*) starts with the definition of GLM and two derivative models (normal linear and binary response regression models). Logit and probit models as instances of binary response regression models are treated in detail and in depth. Both logit and probit models are explained, exemplified and coded very clearly. Latent variable modeling is also treated extensively with adequate conceptual and mathematical treatment, along with MATLAB code implementing a probit latent variable modeling through the Gibbs sampler. Poisson regression for count data ends this very interesting and practical chapter. Although this chapter is very extensive on presentation, it is rather short in terms of end-of-chapter problems, given its practical nature.

In Chapter 10 (*Dependent Data Models*), the authors present the case where response data are not independent due to serial and temporal dependence and multilevel nesting. Autoregressive models and moving average models, separately and combined, are introduced with motivating examples from marketing and economics contexts. The authors adopt a more integrative approach in this chapter by walking the audience through a concrete example, mathematical modeling, computational techniques and code, output interpretation and the use of graphs. This integration of parts more accurately demonstrates the *computational* aspect of statistical analysis than providing code as a facilitative aid. Multivariate Gaussian models are also introduced and elaborated through Gaussian graphical models with an example. The authors next proceed to random effects and mixed effects ANOVA models as other instances of dependent observations. This last section extensively uses code and numerical methods for parameter estimation, which once again shows how skillfully modeling and computing are integrated. This chapter ends with both mathematical and computational problems, but again they are noticeably fewer than the problems in Parts I and II of the book.

Chapter 11 (*State Space Models*) builds on the dependent observations data introduced in the previous chapter adding high-dimensional concepts from stochastic and dynamical systems (or control engineering) fields where model parameters depend on time. The authors first introduce linear Gaussian discrete-time state space models (using Kalman filter). The first model introduced is the unobserved components model. The authors use inflation data as a working example and MATLAB-coded MLE to walk the reader through the concept and its implementation with rigorous formalism (Bayesian estimation is also explained for this model). Time-varying parameter model is the next state space model presented by the authors, recast as a linear regression with nonconstant coefficients. The model is further elaborated using Bayesian estimation methods and exemplified using the same inflation data with MATLAB code. Stochastic volatility model, widely used in economic and financial data, is introduced next as a nonlinear state space model. Given its non-linearity challenge, the authors present an auxiliary mixture sampling approach as the model parameter estimation method, through a worked example and MATLAB computation code. This chapter ends with a brief set of mainly mathematically-oriented problems.

The book ends with two appendices. The first appendix is an 18-page overview of the MATLAB program, syntax, functions and graphs, which is fairly adequate for the unfamiliar reader to understand the code in the main text. The second appendix includes three pages on multivariate differentiation and proofs of some of the theorems in the main text. Finally, solutions are provided for select end-of-chapter problems.

The authors in *Statistical Modeling and Computation* have attempted to make the point that statistical thinking can be cultivated in students through concomitant presentation and integration of theory, mathematics, and coded algorithms. This is clearly stated in the preface to the book and implemented in all sections of the book: "Throughout the book our leading principle is that the mathematical formulation of a statistical model goes hand in hand with the specification of its simulation counterpart" (p. viii). The chapters in Parts I and II prepare the students with the theory and mathematical statistics needed to understand the principles of statistical modeling and the chapters in Part III present the reader with advanced applications of modeling in statistical fields. In doing so, the book benefits from some strengths and is limited by some drawbacks.

A prominent strength of the book is the use and integration of a programming language (MATLAB) in presenting and implementing statistical models, with a heavier emphasis on numerical methods. Coding is gradually introduced from earlier chapters and reaches a central position in chapters in Part III, which finely integrate modeling and computation with real-life examples and data. Although the reader may perceive computation in the earlier chapters (especially in Part I) to be pedagogical aid, they will understand the central role of computation in modern statistics from Chapter 6 onwards and specifically in Part III of the book. The authors provide written functions for those who may not have the **Statistical Toolbox** in MATLAB.

The choice of a command-line language instead of point-and-click or applets (as adopted in some undergraduate textbooks) is a great feature of the book because unlike black-box methods, command-line tools directly map much of the statistical and mathematical models and formulas and provide the user with freedom for manipulation and research.

In terms of statistical concepts, the book is a standard text which includes both classical and Bayesian approaches to modeling and data analysis. Fundamentals of probability and modeling are presented in a rigorous language and the transition to more advanced chapters is almost smooth. Explanations are precise, both verbally and mathematically. Throughout the book, cross-references are made so that the reader can find further or related topics in other parts of the book. Readers interested in mathematical rigor will find this book rewarding. In addition, most exercises involve mathematical derivations and proofs. Therefore, this text could be used for an undergraduate course in probability and statistics.

Another strength of the book lies in the wealth and variety of exercises at the end of each chapter. The exercises (some with complete solutions) range from mathematical proofs and model building to programming. Solutions for select problems are presented at the end of the book.

While the book is outstanding in terms of coverage of topics, rigorous language and integration of computation, I find some aspects of the text slightly less appealing, especially for the novice.

Pedagogically, the reader would benefit more if concepts were presented with more elaborate exposition. Verbal exposition is noticeably replaced with mathematical formalism. More elaborate explanations of the concepts and extension on examples would engage the reader more as the primary goal and requirement of statistical modeling is the development of conceptual thinking. Overall, the book is dense and requires patience from both the reader and instructor to make the best of its wealth of information. Furthermore, some common statistical methods could be added to the text, including validation methods, nonparametric statistics, classification methods and data reduction methods as dedicated chapters. Also, topics on regression would benefit from inclusion of extensive model diagnostics.

In addition, the mathematical rigor and language used throughout the book tends to slightly overshadow the *computational* aspects of modeling, especially in the middle chapters of the book. Although the authors begin with the premise that mathematical formulation is tied to simulation, the underemphasis on the conceptual and computational aspects of modeling may make this book seem more of mathematical statistics than a work that tries to teach modeling primarily.

And lastly, although equivalent R code is provided on the companion website, emphasis on R as the programming language in the book (and with a short primer) would familiarize the readers with the power of this more common *statistical* language. According to some online surveys, R is becoming more widely used in statistical community than most other software and more jobs in the market are demanding R (along with some commercial or free packages such as SAS, Stata, and SPSS or Python) as a required skill for applicants. Therefore, introducing R as a more common language in the industry and academia, with high expandability and rich package repertoire would benefit the readers as a valuable skill. I believe that each generation should be raised on the best programing language of their era which is also future-proof.

On balance, the book makes a good choice for advanced undergraduate students and first year graduate students in statistical sciences or any independent reader who is inclined to the mathematics of modeling and computation. However, as stated above, the text is limited in expository language and explanations. Therefore, novice learners who primarily seek a smoother entry into statical modeling and computation may find this book daunting (a better choice for the beginners could be Martinez and Martinez 2007). However, doctoral students and researchers from quantitative fields, such as economics, biostatistics, management, psychometrics and physical sciences, who may need an organized treatment and foundation for their work but who already have had exposure to topics in statistics will benefit from the book as well. This book can also make a good text for non-statistics graduate programs requiring an introduction to probability and statistics and modeling.

References

Martinez WL, Martinez AR (2007). *Computational Statistics Handbook with MATLAB*. Chapman & Hall/CRC, Boca Raton.

Reviewer:

Abdolvahab Khademi
University of Massachusetts
College of Education &
Department of Mathematics and Statistics
Amherst MA 01002, United States of America
E-mail: akhademisham@umass.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 66, Book Review 3 *Published:* 2015-08-06
August 2015