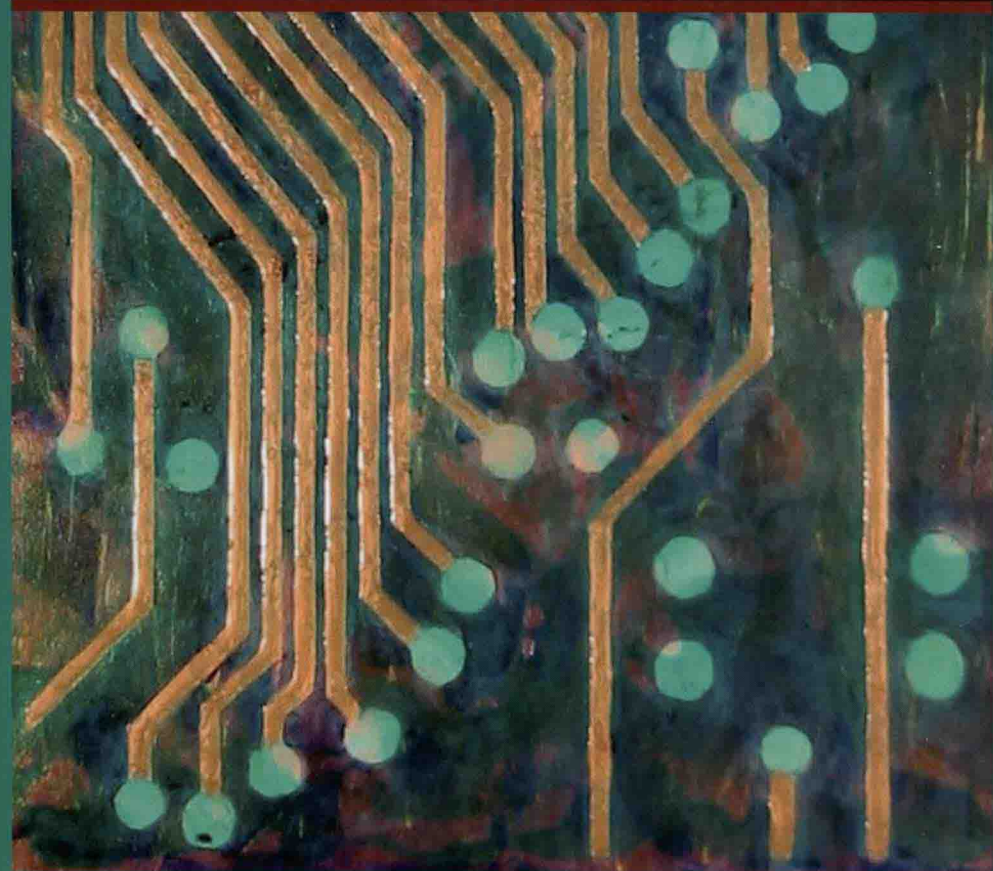


Wiley Series in Computational Statistics

Computational Statistics

Second Edition

GEOF H. GIVENS
JENNIFER A. HOETING

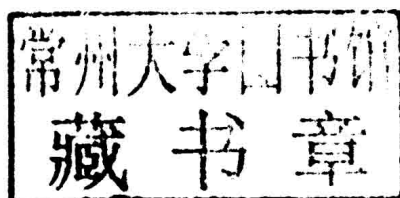


SECOND EDITION

COMPUTATIONAL STATISTICS

GEOF H. GIVENS AND JENNIFER A. HOETING

Department of Statistics, Colorado State University, Fort Collins, CO



 **WILEY**

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2013 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Givens, Geof H.

Computational statistics / Geof H. Givens, Jennifer A. Hoeting. – 2nd ed.
p. cm.

Includes index.

ISBN 978-0-470-53331-4 (cloth)

1. Mathematical statistics—Data processing. I. Hoeting, Jennifer A.
(Jennifer Ann), 1966— II. Title.

QA276.4.G58 2013

519.5—dc23

2012017381

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

COMPUTATIONAL STATISTICS

Wiley Series in Computational Statistics

Consulting Editors:

Paolo Giudici
University of Pavia, Italy

Geof H. Givens
Colorado State University, USA

Bani K. Mallick
Texas A&M University, USA

Wiley Series in Computational Statistics is comprised of practical guides and cutting edge research books on new developments in computational statistics. It features quality authors with a strong applications focus. The texts in the series provide detailed coverage of statistical concepts, methods, and case studies in areas at the interface of statistics, computing, and numerics.

With sound motivation and a wealth of practical examples, the books show in concrete terms how to select and to use appropriate ranges of statistical computing techniques in particular fields of study. Readers are assumed to have a basic understanding of introductory terminology.

The series concentrates on applications of computational methods in statistics to fields of bioinformatics, genomics, epidemiology, business, engineering, finance, and applied statistics.

A complete list of titles in this series appears at the end of the volume

To Natalie and Neil

PREFACE

This book covers most topics needed to develop a broad and thorough working knowledge of modern computational statistics. We seek to develop a practical understanding of how and why existing methods work, enabling readers to use modern statistical methods effectively. Since many new methods are built from components of existing techniques, our ultimate goal is to provide scientists with the tools they need to contribute new ideas to the field.

A growing challenge in science is that there is so much of it. While the pursuit of important new methods and the honing of existing approaches is a worthy goal, there is also a need to organize and distill the teeming jungle of ideas. We attempt to do that here. Our choice of topics reflects our view of what constitutes the core of the evolving field of computational statistics, and what will be interesting and useful for our readers.

Our use of the adjective *modern* in the first sentence of this preface is potentially troublesome: There is no way that this book can cover all the latest, greatest techniques. We have not even tried. We have instead aimed to provide a reasonably up-to-date survey of a broad portion of the field, while leaving room for diversions and esoterica.

The foundations of optimization and numerical integration are covered in this book. We include these venerable topics because (i) they are cornerstones of frequentist and Bayesian inference; (ii) routine application of available software often fails for hard problems; and (iii) the methods themselves are often secondary components of other statistical computing algorithms. Some topics we have omitted represent important areas of past and present research in the field, but their priority here is lowered by the availability of high-quality software. For example, the generation of pseudo-random numbers is a classic topic, but one that we prefer to address by giving students reliable software. Finally, some topics (e.g., principal curves and tabu search) are included simply because they are interesting and provide very different perspectives on familiar problems. Perhaps a future researcher may draw ideas from such topics to design a creative and effective new algorithm.

In this second edition, we have both updated and broadened our coverage, and we now provide computer code. For example, we have added new MCMC topics to reflect continued activity in that popular area. A notable increase in breadth is our inclusion of more methods relevant for problems where statistical dependency is important, such as block bootstrapping and sequential importance sampling. This second edition provides extensive new support in **R**. Specifically, code for the examples in this book is available from the book website www.stat.colostate.edu/computationalstatistics.

Our target audience includes graduate students in statistics and related fields, statisticians, and quantitative empirical scientists in other fields. We hope such readers may use the book when applying standard methods and developing new methods.

The level of mathematics expected of the reader does not extend much beyond Taylor series and linear algebra. ~~Breadth of mathematical training is more helpful than depth.~~ Essential review is provided in Chapter 1. More advanced readers will find greater mathematical detail in the wide variety of high-quality books available on specific topics, many of which are referenced in the text. Other readers caring less about analytical details may prefer to focus on our descriptions of algorithms and examples.

The expected level of statistics is equivalent to that obtained by a graduate student in his or her first year of study of the theory of statistics and probability. An understanding of maximum likelihood methods, Bayesian methods, elementary asymptotic theory, Markov chains, and linear models is most important. Many of these topics are reviewed in Chapter 1.

With respect to computer programming, we find that good students can learn as they go. However, a working knowledge of a suitable language allows implementation of the ideas covered in this book to progress much more quickly. We have chosen to forgo any language-specific examples, algorithms, or coding in the text. For those wishing to learn a language while they study this book, we recommend that you choose a high-level, interactive package that permits the flexible design of graphical displays and includes supporting statistics and probability functions, such as **R** and MATLAB.¹ These are the sort of languages often used by researchers during the development of new statistical computing techniques, and they are suitable for implementing all the methods we describe, except in some cases for problems of vast scope or complexity. We use **R** and recommend it. Although lower-level languages such as C++ could also be used, they are more appropriate for professional-grade implementation of algorithms after researchers have refined the methodology.

The book is organized into four major parts: optimization (Chapters 2, 3, and 4), integration and simulation (Chapters 5, 6, 7, and 8), bootstrapping (Chapter 9) and density estimation and smoothing (Chapters 10, 11, and 12). The chapters are written to stand independently, so a course can be built by selecting the topics one wishes to teach. For a one-semester course, our selection typically weights most heavily topics from Chapters 2, 3, 6, 7, 9, 10, and 11. With a leisurely pace or more thorough coverage, a shorter list of topics could still easily fill a semester course. There is sufficient material here to provide a thorough one-year course of study, notwithstanding any supplemental topics one might wish to teach.

A variety of homework problems are included at the end of each chapter. Some are straightforward, while others require the student to develop a thorough understanding of the model/method being used, to carefully (and perhaps cleverly) code a suitable technique, and to devote considerable attention to the interpretation of results. A few exercises invite open-ended exploration of methods and ideas. We are sometimes asked for solutions to the exercises, but we prefer to sequester them to preserve the challenge for future students and readers.

The datasets discussed in the examples and exercises are available from the book website, www.stat.colostate.edu/computationalstatistics. The **R** code is also provided there. Finally, the website includes an errata. Responsibility for all errors lies with us.

¹**R** is available for free from www.r-project.org. Information about MATLAB can be found at www.mathworks.com.

ACKNOWLEDGMENTS

The course upon which this book is based was developed and taught by us at Colorado State University from 1994 onwards. Thanks are due to our many students who have been semiwilling guinea pigs over the years. We also thank our colleagues in the Statistics Department for their continued support. The late Richard Tweedie merits particular acknowledgment for his mentoring during the early years of our careers.

We owe a great deal of intellectual debt to Adrian Raftery, who deserves special thanks not only for his teaching and advising, but also for his unwavering support and his seemingly inexhaustible supply of good ideas. In addition, we thank our influential advisors and teachers at the University of Washington Statistics Department, including David Madigan, Werner Stuetzle, and Judy Zeh. Of course, each of our chapters could be expanded into a full-length book, and great scholars have already done so. We owe much to their efforts, upon which we relied when developing our course and our manuscript.

Portions of the first edition were written at the Department of Mathematics and Statistics, University of Otago, in Dunedin, New Zealand, whose faculty we thank for graciously hosting us during our sabbatical in 2003. Much of our work on the second edition was undertaken during our sabbatical visit to the Australia Commonwealth Scientific and Research Organization in 2009–2010, sponsored by CSIRO Mathematics, Informatics and Statistics, and hosted at the Longpocket Laboratory in Indooroopilly, Australia. We thank our hosts and colleagues there for their support.

Our manuscript has been greatly improved through the constructive reviews of John Bickham, Ben Bird, Kate Cowles, Jan Hannig, Alan Herlihy, David Hunter, Devin Johnson, Michael Newton, Doug Nychka, Steve Sain, David W. Scott, N. Scott Urquhart, Haonan Wang, Darrell Whitley, and eight anonymous referees. We also thank the sharp-eyed readers listed in the errata for their suggestions and corrections. Our editor Steve Quigley and the folks at Wiley were supportive and helpful during the publication process. We thank Nélida Pohl for permission to adapt her photograph in the cover design of the first edition. We thank Melinda Stelzer for permission to use her painting “Champagne Circuit,” 2001, for the cover of the second edition. More about her art can be found at www.facebook.com/geekchicart. We also owe special note of thanks to Zube (a.k.a. John Dzuber), who kept our own computers running despite our best efforts to the contrary.

Funding from National Science Foundation (NSF) CAREER grant #SBR-9875508 was a significant source of support for the first author during the preparation of the first edition. He also thanks his colleagues and friends in the North Slope Borough, Alaska, Department of Wildlife Management for their longtime research support. The second author gratefully acknowledges the support of STAR Research

Assistance Agreement CR-829095 awarded to Colorado State University by the U.S. Environmental Protection Agency (EPA). The views expressed here are solely those of the authors. NSF and EPA do not endorse any products or commercial services mentioned herein.

Finally, we thank our parents for enabling and supporting our educations and for providing us with the “stubbornness gene” necessary for graduate school, the tenure track, or book publication—take your pick! The second edition is dedicated to our kids, Natalie and Neil, for continuing to show us what is important and what is not.

Geof H. Givens
Jennifer A. Hoeting

CONTENTS

<i>PREFACE</i>	xv
----------------	----

<i>ACKNOWLEDGMENTS</i>	xvii
------------------------	------

xi <i>REVIEW</i>	1
1.1 Mathematical Notation	1
1.2 Taylor's Theorem and Mathematical Limit Theory ✓	2
1.3 Statistical Notation and Probability Distributions ✓	4
1.4 Likelihood Inference ✓	9
1.5 Bayesian Inference ✓	11
1.6 Statistical Limit Theory ✓	13
1.7 Markov Chains ✓	14
1.8 Computing ✓	17

PART I

OPTIMIZATION

2 <i>OPTIMIZATION AND SOLVING NONLINEAR EQUATIONS</i>	21
2.1 Univariate Problems	22
2.1.1 Newton's Method	26
2.1.1.1 Convergence Order	29
2.1.2 Fisher Scoring	30
2.1.3 Secant Method	30
2.1.4 Fixed-Point Iteration	32
2.1.4.1 Scaling	33
2.2 Multivariate Problems	34
2.2.1 Newton's Method and Fisher Scoring ✓	34
2.2.1.1 Iteratively Reweighted Least Squares	36
2.2.2 Newton-Like Methods	39
2.2.2.1 Ascent Algorithms	39
2.2.2.2 Discrete Newton and Fixed-Point Methods	41
2.2.2.3 Quasi-Newton Methods	41

2.2.3	Gauss–Newton Method	44
2.2.4	Nelder–Mead Algorithm	45
2.2.5	Nonlinear Gauss–Seidel Iteration	52
	Problems	54

3 COMBINATORIAL OPTIMIZATION 59

3.1	Hard Problems and NP-Completeness	59
3.1.1	Examples	61
3.1.2	Need for Heuristics	64
3.2	Local Search	65
3.3	Simulated Annealing	68
3.3.1	Practical Issues	70
3.3.1.1	Neighborhoods and Proposals	70
3.3.1.2	Cooling Schedule and Convergence	71
3.3.2	Enhancements	74
3.4	Genetic Algorithms	75
3.4.1	Definitions and the Canonical Algorithm	75
3.4.1.1	Basic Definitions	75
3.4.1.2	Selection Mechanisms and Genetic Operators	76
3.4.1.3	Allele Alphabets and Genotypic Representation	78
3.4.1.4	Initialization, Termination, and Parameter Values	79
3.4.2	Variations	80
3.4.2.1	Fitness	80
3.4.2.2	Selection Mechanisms and Updating Generations	81
3.4.2.3	Genetic Operators and Permutation Chromosomes	82
3.4.3	Initialization and Parameter Values	84
3.4.4	Convergence	84
3.5	Tabu Algorithms	85
3.5.1	Basic Definitions	86
3.5.2	The Tabu List	87
3.5.3	Aspiration Criteria	88
3.5.4	Diversification	89
3.5.5	Intensification	90
3.5.6	Comprehensive Tabu Algorithm	91
	Problems	92


4 EM OPTIMIZATION METHODS 97

4.1	Missing Data, Marginalization, and Notation	97
4.2	The EM Algorithm	98
4.2.1	Convergence	102
4.2.2	Usage in Exponential Families	105

4.2.3	Variance Estimation	106
4.2.3.1	Louis's Method	106
4.2.3.2	SEM Algorithm	108
4.2.3.3	Bootstrapping	110
4.2.3.4	Empirical Information	110
4.2.3.5	Numerical Differentiation	111
4.3	EM Variants	111
4.3.1	Improving the E Step	111
4.3.1.1	Monte Carlo EM	111
4.3.2	Improving the M Step	112
4.3.2.1	ECM Algorithm	113
4.3.2.2	EM Gradient Algorithm	116
4.3.3	Acceleration Methods	118
4.3.3.1	Aitken Acceleration	118
4.3.3.2	Quasi-Newton Acceleration	119
	Problems	121

PART II

INTEGRATION AND SIMULATION

5	NUMERICAL INTEGRATION	129
5.1	Newton–Côtes Quadrature	129
5.1.1	Riemann Rule	130
5.1.2	Trapezoidal Rule	134
5.1.3	Simpson's Rule	136
5.1.4	General k th-Degree Rule	138
5.2	Romberg Integration	139
5.3	Gaussian Quadrature	142
5.3.1	Orthogonal Polynomials	143
5.3.2	The Gaussian Quadrature Rule	143
5.4	Frequently Encountered Problems	146
5.4.1	Range of Integration	146
5.4.2	Integrands with Singularities or Other Extreme Behavior	146
5.4.3	Multiple Integrals	147
5.4.4	Adaptive Quadrature	147
5.4.5	Software for Exact Integration	148
	Problems	148
	SIMULATION AND MONTE CARLO INTEGRATION	151
6.1	Introduction to the Monte Carlo Method	151
6.2	Exact Simulation	152

6.2.1	Generating from Standard Parametric Families	153
6.2.2	Inverse Cumulative Distribution Function	153
6.2.3	Rejection Sampling	155
6.2.3.1	Squeezed Rejection Sampling	158
6.2.3.2	Adaptive Rejection Sampling	159
6.3	Approximate Simulation	163
6.3.1	Sampling Importance Resampling Algorithm	163
6.3.1.1	Adaptive Importance, Bridge, and Path Sampling	167
6.3.2	<u>Sequential Monte Carlo</u>	168
6.3.2.1	Sequential Importance Sampling for Markov Processes	169
6.3.2.2	General Sequential Importance Sampling	170
6.3.2.3	Weight Degeneracy, Rejuvenation, and Effective Sample Size	171
6.3.2.4	Sequential Importance Sampling for Hidden Markov Models	175
6.3.2.5	Particle Filters	179
6.4	Variance Reduction Techniques	180
6.4.1	Importance Sampling	180
6.4.2	Antithetic Sampling	186
6.4.3	Control Variates	189
6.4.4	Rao–Blackwellization	193
	Problems	195


MARKOV CHAIN MONTE CARLO 201

7.1	Metropolis–Hastings Algorithm	202
7.1.1	Independence Chains	204
7.1.2	Random Walk Chains	206
7.2	Gibbs Sampling	209
7.2.1	Basic Gibbs Sampler	209
7.2.2	Properties of the Gibbs Sampler	214
7.2.3	Update Ordering	216
7.2.4	Blocking	216
7.2.5	Hybrid Gibbs Sampling	216
7.2.6	Griddy–Gibbs Sampler	218
7.3	Implementation	218
7.3.1	Ensuring Good Mixing and Convergence	219
7.3.1.1	Simple Graphical Diagnostics	219
7.3.1.2	Burn-in and Run Length	220
7.3.1.3	Choice of Proposal	222
7.3.1.4	Reparameterization	223
7.3.1.5	Comparing Chains: Effective Sample Size	224
7.3.1.6	Number of Chains	225

7.3.2	Practical Implementation Advice	226
7.3.3	Using the Results	226
	Problems	230
8	ADVANCED TOPICS IN MCMC	237
8.1	Adaptive MCMC	237
8.1.1	Adaptive Random Walk Metropolis-within-Gibbs Algorithm	238
8.1.2	General Adaptive Metropolis-within-Gibbs Algorithm	240
8.1.3	Adaptive Metropolis Algorithm	247
8.2	Reversible Jump MCMC	250
8.2.1	RJMCMC for Variable Selection in Regression	253
8.3	Auxiliary Variable Methods	256
8.3.1	Simulated Tempering	257
8.3.2	Slice Sampler	258
8.4	Other Metropolis–Hastings Algorithms	260
8.4.1	Hit-and-Run Algorithm	260
8.4.2	Multiple-Try Metropolis–Hastings Algorithm	261
8.4.3	Langevin Metropolis–Hastings Algorithm	262
8.5	Perfect Sampling	264
8.5.1	Coupling from the Past	264
8.5.1.1	Stochastic Monotonicity and Sandwiching	267
8.6	Markov Chain Maximum Likelihood	268
8.7	Example: MCMC for Markov Random Fields	269
8.7.1	Gibbs Sampling for Markov Random Fields	270
8.7.2	Auxiliary Variable Methods for Markov Random Fields	274
8.7.3	Perfect Sampling for Markov Random Fields	277
	Problems	279

PART III



BOOTSTRAPPING

	BOOTSTRAPPING	287
9.1	The Bootstrap Principle	287
9.2	Basic Methods	288
9.2.1	Nonparametric Bootstrap	288
9.2.2	Parametric Bootstrap	289
9.2.3	Bootstrapping Regression	290
9.2.4	Bootstrap Bias Correction	291
9.3	Bootstrap Inference	292
9.3.1	Percentile Method	292
9.3.1.1	Justification for the Percentile Method	293
9.3.2	Pivoting	294

9.3.2.1	Accelerated Bias-Corrected Percentile Method, BC_a	294
9.3.2.2	The Bootstrap t	296
9.3.2.3	Empirical Variance Stabilization	298
9.3.2.4	Nested Bootstrap and Prepivoting	299
9.3.3	Hypothesis Testing	301
9.4	Reducing Monte Carlo Error	302
9.4.1	Balanced Bootstrap	302
9.4.2	Antithetic Bootstrap	302
9.5	Bootstrapping Dependent Data	303
9.5.1	Model-Based Approach	304
9.5.2	Block Bootstrap	304
9.5.2.1	Nonmoving Block Bootstrap	304
9.5.2.2	Moving Block Bootstrap	306
9.5.2.3	Blocks-of-Blocks Bootstrapping	307
9.5.2.4	Centering and Studentizing	309
9.5.2.5	Block Size	311
9.6	Bootstrap Performance	315
9.6.1	Independent Data Case	315
9.6.2	Dependent Data Case	316
9.7	Other Uses of the Bootstrap	316
9.8	Permutation Tests	317
	Problems	319

PART IV

DENSITY ESTIMATION AND SMOOTHING

		NONPARAMETRIC DENSITY ESTIMATION	325
10.1	Measures of Performance		326
10.2	Kernel Density Estimation		327
10.2.1	Choice of Bandwidth		329
10.2.1.1	Cross-Validation		332
10.2.1.2	Plug-in Methods		335
10.2.1.3	Maximal Smoothing Principle		338
10.2.2	Choice of Kernel		339
10.2.2.1	Epanechnikov Kernel		339
10.2.2.2	Canonical Kernels and Rescalings		340
10.3	Nonkernel Methods		341
10.3.1	Logspline		341
10.4	Multivariate Methods		345
10.4.1	The Nature of the Problem		345