Computational Statistics

First Edition

Geof H. Givens and Jennifer A. Hoeting

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

New York / Chichester / Weinheim / Brisbane / Singapore / Toronto

Preface

This book covers most topics needed to develop a broad and thorough working knowledge of modern statistical computing and 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.

Achieving these goals requires familiarity with diverse topics in statistical computing, computational statistics, computer science, and numerical analysis. Our choice of topics reflects our view of what is central to this evolving field, and what will be interesting and useful for our readers. We pragmatically assigned priority to topics that can be of the most benefit to students and researchers most quickly.

Some topics we 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. Some topics, such as numerical linear algebra, are on the borderline. Such topics are critical for many applications, yet good routines are generally available. In our judgment, the frequency with which one must shelve the routines and dig into the details of numerical linear algebra falls (barely) below the threshold we set for inclusion in this book. Among the classic topics we have chosen to cover are optimization and numerical integration. We include these 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.

Our use of the adjective *modern* is potentially troublesome: there is no way that this book can cover all the latest, greatest techniques. We have not even tried. Some topics, such as heuristic search and Markov chain Monte Carlo, simply move too quickly. We have instead tried to offer a reasonably up-to-date survey of a broad portion of the field, while leaving room for diversions and esoterica. 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.

Our target audience includes graduate students in statistics and related fields, working 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. For those wishing to learn a language while they study this book, we recommend you choose a high-level, interactive package that permits the flexible design of graphical displays and includes supporting statistics and probability functions. At the time of writing, we recommend S-Plus, R, and MATLAB.[1] These are the sort of languages often used by researchers during the development of new statistical computing techniques, and are suitable for implementing all the methods we describe, except in some cases for problems of vast scope or complexity. Of course, lower-level languages such as C++ can also be used, and are favored for professional grade implementation of algorithms after researchers have refined the methodology.

Even adept computer programmers may have little understanding of how mathematics is carried out in the binary world of a computer. Mysterious problems with

[1] Websites for these software packages are www.insightful.com, www.r-project.org, and www.mathworks.com, respectively. R is free software reimplementing portions of S-Plus; the others are commercial.

full-rank matrices that appear noninvertible, integrals and likelihoods that vanish, numerical approximations that appear more precise than they really are, and other oddities are not unusual. While not dismissing the importance of computer arithmetic and numerically stable computation, we prefer to focus on the big picture of how algorithms work and to sweep under the rug some of the nitty-gritty numerical computation details.

The book is organized into three major parts: optimization (Chapters 2, 3, and 4), integration (Chapters 5, 6, 7, and 8), and smoothing (Chapters 10, 11, and 12). Chapter 9 adds another essential topic, the bootstrap. 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, 5, 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.

The datasets discussed in the text and problems are available from the book website, *www.stat.colostate.edu/computationalstatistics*. The errata will also be found there. Responsibility for all errors lies with us.

Geof H. Givens and Jennifer A. Hoeting

Fort Collins, Colorado

Contents

Preface **vii**

Acknowledgments **xi**

1 Review **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*

2 Optimization and Solving Nonlinear Equations **19**

2.1 Univariate problems *20*

2.1.1 Newton's method *24*

2.1.1.1 Convergence order *27*

2.1.2 Fisher scoring *28*

2.1.3 Secant method *28*

2.1.4 Fixed-point iteration *30*

2.1.4.1 Scaling *31*

2.2 Multivariate problems *31*

2.2.1 Newton's method and Fisher scoring 32
2.2.1.1 Iteratively reweighted least squares 34
2.2.2 Newton-like methods 37
2.2.2.1 Ascent algorithms 37
2.2.2.2 Discrete Newton and fixed-point methods 39
2.2.2.3 Quasi-Newton methods 39
2.2.3 Gauss–Newton method 42
2.2.4 Nonlinear Gauss–Seidel iteration and other methods 43
Problems 44

3 Combinatorial Optimization 49
3.1 Hard problems and NP-completeness 50
3.1.1 Examples 51
3.1.2 The need for heuristics 55
3.2 Local search 55
3.3 Tabu algorithms 59
3.3.1 Basic definitions 60
3.3.2 The tabu list 61
3.3.3 Aspiration criteria 63
3.3.4 Diversification 64
3.3.5 Intensification 65
3.3.6 A comprehensive tabu algorithm 65
3.4 Simulated annealing 67
3.4.1 Practical issues 68
3.4.1.1 Neighborhoods and proposals 68
3.4.1.2 Cooling schedule and convergence 69
3.4.2 Enhancements 72
3.5 Genetic algorithms 73
3.5.1 Definitions and the canonical algorithm 74
3.5.1.1 Basic definitions 74
3.5.1.2 Selection mechanisms and genetic operators 75
3.5.1.3 Allele alphabets and genotypic representation 76
3.5.1.4 Initialization, termination, and parameter values 78
3.5.2 Variations 78
3.5.2.1 Fitness 78
3.5.2.2 Selection mechanisms and updating generations 79

3.5.2.3 Genetic operators and permutation chromosomes 80
3.5.3 Initialization and parameter values 83
3.5.4 Convergence 83
Problems 84

4 EM Optimization Methods 89
4.1 Missing data, marginalization, and notation 90
4.2 The EM algorithm 90
4.2.1 Convergence 95
4.2.2 Usage in exponential families 97
4.2.3 Variance estimation 98
4.2.3.1 Louis's method 99
4.2.3.2 SEM algorithm 101
4.2.3.3 Bootstrapping 102
4.2.3.4 Empirical information 103
4.2.3.5 Numerical differentiation 103
4.3 EM Variants 104
4.3.1 Improving the E step 104
4.3.1.1 Monte Carlo EM 104
4.3.2 Improving the M step 105
4.3.2.1 ECM algorithm 105
4.3.2.2 EM gradient algorithm 109
4.3.3 Acceleration methods 110
4.3.3.1 Aitken acceleration 110
4.3.3.2 Quasi-Newton acceleration 111
Problems 113

5 Numerical Integration 121
5.1 Newton–Côtes quadrature 122
5.1.1 Riemann rule 122
5.1.2 Trapezoidal rule 126
5.1.3 Simpson's rule 129
5.1.4 General kth-degree rule 131
5.2 Romberg integration 131
5.3 Gaussian quadrature 135
5.3.1 Orthogonal polynomials 136
5.3.2 The Gaussian quadrature rule 136
5.4 Frequently encountered problems 139
5.4.1 Range of integration 139

5.4.2 Integrands with singularities or other extreme behavior 139
5.4.3 Multiple integrals 140
5.4.4 Adaptive quadrature 140
5.4.5 Software for exact integration 140
Problems 141

6 Simulation and Monte Carlo Integration 143
6.1 Introduction to the Monte Carlo method 144
6.2 Simulation 145
6.2.1 Generating from standard parametric families 145
6.2.2 Inverse cumulative distribution function 145
6.2.3 Rejection sampling 147
6.2.3.1 Squeezed rejection sampling 150
6.2.3.2 Adaptive rejection sampling 151
6.2.4 The sampling importance resampling algorithm 155
6.2.4.1 Adaptive importance, bridge, and path sampling 159
6.2.4.2 Sequential importance sampling 161
6.3 Variance reduction techniques 162
6.3.1 Importance sampling 163
6.3.2 Antithetic sampling 169
6.3.3 Control variates 172
6.3.4 Rao–Blackwellization 176
Problems 177

7 Markov Chain Monte Carlo 183
7.1 Metropolis–Hastings algorithm 184
7.1.1 Independence chains 186
7.1.2 Random walk chains 188
7.1.3 Hit-and-run algorithm 191
7.1.4 Langevin Metropolis–Hastings algorithm 193
7.1.5 Multiple-try Metropolis–Hastings algorithm 194
7.2 Gibbs sampling 195
7.2.1 Basic Gibbs sampler 195
7.2.2 Immediate updating 198
7.2.3 Update ordering 198
7.2.4 Blocking 199
7.2.5 Hybrid Gibbs sampling 199
7.2.6 Alternative univariate proposal methods 200

7.3 Implementation 200
7.3.1 Ensuring good mixing and convergence 201
7.3.1.1 Choice of proposal 201
7.3.1.2 Number of chains 202
7.3.1.3 Simple graphs to assess mixing and convergence 203
7.3.1.4 Reparameterization 204
7.3.1.5 Burn-in and run length 205
7.3.2 Practical implementation advice 206
7.3.3 Using the results 207
7.4 Example: Fur seal pup capture–recapture data 208
Problems 212

8 Advanced Topics in MCMC 219
8.1 Auxiliary variable methods 219
8.1.1 Slice sampler 221
8.2 Reversible Jump MCMC 224
8.2.1 RJMCMC for variable selection in regression 227
8.3 Perfect sampling 230
8.3.1 Coupling from the past 231
8.3.1.1 Stochastic monotonicity and sandwiching 234
8.4 Example: MCMC for Markov random fields 235
8.4.1 Gibbs sampling for Markov random fields 236
8.4.2 Auxiliary variable methods for Markov random fields 241
8.4.3 Perfect sampling for Markov random fields 244
8.5 Markov chain maximum likelihood 246
Problems 247

9 Bootstrapping 253
9.1 The bootstrap principle 253
9.2 Basic methods 255
9.2.1 Nonparametric bootstrap 255
9.2.2 Parametric bootstrap 256
9.2.3 Bootstrapping regression 256
9.2.4 Bootstrap bias correction 258
9.3 Bootstrap inference 259
9.3.1 Percentile method 259
9.3.1.1 Justification for the percentile method 260
9.3.2 Pivoting 261

9.3.2.1 Accelerated bias-corrected percentile method, BC_a 261
9.3.2.2 The bootstrap t 263
9.3.2.3 Empirical variance stabilization 264
9.3.2.4 Nested bootstrap and prepivoting 265
9.3.3 Hypothesis testing 268
9.4 Reducing Monte Carlo error 268
9.4.1 Balanced bootstrap 268
9.4.2 Antithetic bootstrap 269
9.5 Other uses of the bootstrap 269
9.6 Degree of bootstrap approximation 270
9.7 Permutation tests 272
Problems 273

10 Nonparametric Density Estimation 277
10.1 Measures of performance 278
10.2 Kernel density estimation 280
10.2.1 Choice of bandwidth 281
10.2.1.1 Cross-validation 284
10.2.1.2 Plug-in methods 288
10.2.1.3 Maximal smoothing principle 290
10.2.2 Choice of kernel 292
10.2.2.1 Epanechnikov kernel 292
10.2.2.2 Canonical kernels and rescalings 293
10.3 Nonkernel methods 294
10.3.1 Logspline 294
10.4 Multivariate methods 297
10.4.1 The nature of the problem 298
10.4.2 Multivariate kernel estimators 298
10.4.3 Adaptive kernels and nearest neighbors 301
10.4.3.1 Nearest neighbor approaches 302
10.4.3.2 Variable-kernel approaches and transformations 303
10.4.4 Exploratory projection pursuit 306
Problems 312

11 Bivariate Smoothing 315
11.1 Predictor–response data 316
11.2 Linear smoothers 318
11.2.1 Constant-span running mean 318

11.2.1.1 Effect of span 320
11.2.1.2 Span selection for linear smoothers 322
11.2.2 Running lines and running polynomials 325
11.2.3 Kernel smoothers 326
11.2.4 Local regression smoothing 327
11.2.5 Spline smoothing 329
11.2.5.1 Choice of penalty 330
11.3 Comparison of linear smoothers 330
11.4 Nonlinear smoothers 332
11.4.1 Loess 332
11.4.2 Supersmoother 333
11.5 Confidence bands 337
11.6 General bivariate data 341
Problems 342

12 Multivariate Smoothing 347
12.1 Predictor–response data 347
12.1.1 Additive models 348
12.1.2 Generalized additive models 351
12.1.3 Other methods related to additive models 355
12.1.3.1 Projection pursuit regression 355
12.1.3.2 Neural networks 358
12.1.3.3 Alternating conditional expectations 358
12.1.3.4 Additivity and variance stabilization 359
12.1.4 Tree-based methods 360
12.1.4.1 Recursive partitioning regression trees 362
12.1.4.2 Tree pruning 364
12.1.4.3 Classification trees 367
12.1.4.4 Other issues for tree-based methods 368
12.2 General multivariate data 369
12.2.1 Principal curves 369
12.2.1.1 Definition and motivation 369
12.2.1.2 Estimation 371
12.2.1.3 Span selection 372
Problems 372

Data Acknowledgments 377

References 379

Index 407