Bayesian Data Analysis
Second edition
Corrected version (30 Jan 2008)

Andrew Gelman
Columbia University

John B. Carlin
University of Melbourne, Australia

Hal S. Stern
University of California, Irvine

Donald B. Rubin
Harvard University

Contents

List of models

List of examples

Preface

This book is intended to have three roles and to serve three associated audiences: an introductory text on Bayesian inference starting from first principles, a graduate text on effective current approaches to Bayesian modeling and computation in statistics and related fields, and a handbook of Bayesian methods in applied statistics for general users of and researchers in applied statistics. Although introductory in its early sections, the book is definitely not elementary in the sense of a first text in statistics. The mathematics used in our book is basic probability and statistics, elementary calculus, and linear algebra. A review of probability notation is given in Chapter 1 along with a more detailed list of topics assumed to have been studied. The practical orientation of the book means that the reader's previous experience in probability, statistics, and linear algebra should ideally have included strong computational components.

To write an introductory text alone would leave many readers with only a taste of the conceptual elements but no guidance for venturing into genuine practical applications, beyond those where Bayesian methods agree essentially with standard non-Bayesian analyses. On the other hand, given the continuing scarcity of introductions to applied Bayesian statistics either in books or in statistical education, we feel it would be a mistake to present the advanced methods without first introducing the basic concepts from our data-analytic perspective. Furthermore, due to the nature of applied statistics, a text on current Bayesian methodology would be incomplete without a variety of worked examples drawn from real applications. To avoid cluttering the main narrative, *there are bibliographic notes at the end of each chapter* and references at the end of the book.

Examples of real statistical analyses are found throughout the book, and we hope thereby to give a genuine applied flavor to the entire development. Indeed, given the conceptual simplicity of the Bayesian approach, it is only in the intricacy of specific applications that novelty arises. Non-Bayesian approaches to inference have dominated statistical theory and practice for most of the past century, but the last two decades or so have seen a reemergence of the Bayesian approach. This has been driven more by the availability of new computational techniques than by what many would see as the philosophical and logical advantages of Bayesian thinking.

We hope that the publication of this book will enhance the spread of ideas that are currently trickling through the scientific literature. The models and methods developed recently in this field have yet to reach their largest possible audience, partly because the results are scattered in various journals and proceedings volumes. We hope that this book will help a new generation of statisticians and users of statistics to solve complicated problems with greater understanding.

Progress in Bayesian data analysis

Bayesian methods have matured and improved in several ways in the eight years since the first edition of this book appeared.

- Successful applications of Bayesian data analysis have appeared in many different fields, including business, computer science, economics, educational research, environmental science, epidemiology, genetics, geography, imaging, law, medicine, political science, psychometrics, public policy, sociology, and sports. In the social sciences, Bayesian ideas often appear in the context of multilevel modeling.
- New computational methods generalizing the Gibbs sampler and Metropolis algorithm, including some methods from the physics literature, have been adapted to statistical problems. Along with improvements in computing speed, these have made it possible to compute Bayesian inference for more complicated models on larger datasets.
- In parallel with the theoretical improvements in computation, the software package `Bugs` has allowed nonexperts in statistics to fit complex Bayesian models with minimal programming. Hands-on experience has convinced many applied researchers of the benefits of the Bayesian approach.
- There has been much work on model checking and comparison, from many perspectives, including predictive checking, cross-validation, Bayes factors, model averaging, and estimates of predictive errors and model complexity.
- In sample surveys and elsewhere, multiple imputation has become a standard method of capturing uncertainty about missing data. This has motivated ongoing work into more flexible models for multivariate distributions.
- There has been continuing progress by various researchers in combining Bayesian inference with existing statistical approaches from other fields, such as instrumental variables analysis in economics, and with nonparametric methods such as classification trees, splines, and wavelets.
- In general, work in Bayesian statistics now focuses on applications, computations, and models. Philosophical debates, abstract optimality criteria, and asymptotic analyses are fading to the background. It is now possible to do serious applied work in Bayesian inference without the need to debate foundational principles of inference.

Changes for the second edition

The major changes for the second edition of this book are:

- Reorganization and expansion of Chapters 6 and 7 on model checking and data collection;
- Revision of Part III on computation;

- New chapters on nonlinear models and decision analysis;
- An appendix illustrating computation using the statistical packages `R` and `Bugs`,
- New applied examples throughout, including:
  - Census record linkage, a data-based assignment of probability distributions (Section 1.7),
  - Cancer mapping, demonstrating the role of the prior distribution on data with different sample sizes (Section 2.8),
  - Psychological measurement data and the use of graphics in model checking (Section 6.4),
  - Survey of adolescent smoking, to illustrate numerical predictive checks (Section 6.5),
  - Two surveys using cluster sampling (Section 7.4),
  - Experiment of vitamin A intake, with noncompliance to assigned treatment (Section 7.7),
  - Factorial data on internet connect times, summarized using the analysis of variance (Section 15.6),
  - Police stops, modeled with hierarchical Poisson regressions (Section 16.5),
  - State-level opinions from national polls, using hierarchical modeling and poststratification (Section 16.6),
  - Serial dilution assays, as an example of a nonlinear model (Section 20.2),
  - Data from a toxicology experiment, analyzed with a hierarchical nonlinear model (Section 20.3),
  - Pre-election polls, with multiple imputation of missing data (Section 21.2),
  - Incentives for telephone surveys, a meta-analysis for a decision problem (Section 22.2),
  - Medical screening, an example of a decision analysis (Section 22.3),
  - Home radon measurement and remediation decisions, analyzed using a hierarchical model (Section 22.4).

We have added these examples because our readers have told us that one thing they liked about the book was the presentation of realistic problem-solving experiences. As in the first edition, we have included many applications from our own research because we know enough about these examples to convey the specific challenges that arose in moving from substantive goals to probability modeling and, eventually, to substantive conclusions. Also as before, some of the examples are presented schematically and others in more detail.

We changed the computation sections out of recognition that our earlier recommendations were too rigid: Bayesian computation is currently at a stage where there are many reasonable ways to compute any given posterior distribution, and the best approach is not always clear in advance. Thus we have moved to a more pluralistic presentation—we give advice about performing computations from many perspectives, including approximate computation, mode-finding, and simulations, while making clear, especially in the discussion of individual models in the later parts of the book, that it is important to be aware of the different ways of implementing any given iterative simulation computation. We briefly discuss some recent ideas in Bayesian computation but devote most of Part III to the practical issues of implementing the Gibbs sampler and the Metropolis algorithm. Compared to the first edition, we deemphasize approximations based on the normal distribution and the posterior mode, treating these now almost entirely as techniques for obtaining starting points for iterative simulations.

Contents

Part I introduces the fundamental Bayesian principle of treating all unknowns as random variables and presents basic concepts, standard probability models, and some applied examples. In Chapters 1 and 2, simple familiar models using the normal, binomial, and Poisson distributions are used to establish this introductory material, as well as to illustrate concepts such as conjugate and noninformative prior distributions, including an example of a nonconjugate model. Chapter 3 presents the Bayesian approach to multiparameter problems. Chapter 4 introduces large-sample asymptotic results that lead to normal approximations to posterior distributions.

Part II introduces more sophisticated concepts in Bayesian modeling and model checking. Chapter 5 introduces hierarchical models, which reveal the full power and conceptual simplicity of the Bayesian approach for practical problems. We illustrate issues of model construction and computation with a relatively complete Bayesian analysis of an educational experiment and of a meta-analysis of a set of medical studies. Chapter 6 discusses the key practical concerns of model checking, sensitivity analysis, and model comparison, illustrating with several examples. Chapter 7 discusses how Bayesian data analysis is influenced by data collection, including the topics of ignorable and nonignorable data collection rules in sample surveys and designed experiments, and specifically the topic of randomization, which is presented as a device for increasing the robustness of posterior inferences. This a difficult chapter, because it presents important ideas that will be unfamiliar to many readers. Chapter 8 discusses connections to non-Bayesian statistical methods, emphasizing common points in practical applications and current challenges in implementing Bayesian data analysis. Chapter 9 summarizes some of the key ideas of Bayesian modeling, inference, and model checking, illustrating issues with some relatively simple examples that highlight potential pitfalls in trying to fit models automatically.

Part III covers Bayesian computation, which can be viewed as a highly specialized branch of numerical analysis: given a posterior distribution function (possibly implicitly defined), how does one extract summaries such as quantiles, moments, and modes, and draw random samples of values? We emphasize iterative methods—the Gibbs sampler and Metropolis algorithm—for drawing random samples from the posterior distribution.

Part IV discusses regression models, beginning with a Bayesian treatment of classical regression illustrated using an example from the study of elections that has both causal and predictive aspects. The subsequent chapters give general principles and examples of hierarchical linear models, generalized linear models, and robust models.

Part V presents a range of other Bayesian probability models in more detail, with examples of multivariate models, mixtures, and nonlinear models. We conclude with methods for missing data and decision analysis, two practical concerns that arise implicitly or explicitly in many statistical problems.

Throughout, we illustrate in examples the three steps of Bayesian statistics: (1) setting up a full probability model using substantive knowledge, (2) conditioning on observed data to form a posterior inference, and (3) evaluating the fit of the model to substantive knowledge and observed data.

Appendixes provide a list of common distributions with their basic properties, a sketch of a proof of the consistency and limiting normality of Bayesian posterior distributions, and an extended example of Bayesian computation in the statistical packages `Bugs` and `R`.

Most chapters conclude with a set of exercises, including algebraic derivations, simple algebraic and numerical examples, explorations of theoretical topics covered only briefly in the text, computational exercises, and data analyses. The exercises in the later chapters tend to be more difficult; some are suitable for term projects.

One-semester or one-quarter course

This book began as lecture notes for a graduate course. Since then, we have attempted to create an advanced undergraduate text, a graduate text, and a reference work all in one, and so the instructor of any course based on this book must be selective in picking out material.

Chapters 1–6 should be suitable for a one-semester course in Bayesian statistics for advanced undergraduates, although these students might also be interested in the introduction to Markov chain simulation in Chapter 11.

Part I has many examples and algebraic derivations that will be useful for a lecture course for undergraduates but may be left to the graduate students to read at home (or conversely, the lectures can cover the examples and leave the theory for homework). The examples of Part II are crucial, however, since these ideas will be new to most graduate students as well. We see the first two chapters of Part III as essential for understanding modern Bayesian computation and the first three chapters of Part IV as basic to any graduate course because they take the student into the world of standard applied models; the remaining material in Parts III–V can be covered as time permits.

This book has been used as the text for one-semester and one-quarter courses for graduate students in statistics at many universities. We suggest the following syllabus for an intense fifteen-week course.

1. Setting up a probability model, Bayes' rule, posterior means and variances, binomial model, proportion of female births (Chapter 1, Sections 2.1–2.5).
2. Standard univariate models including the normal and Poisson models, cancer rate example, noninformative prior distributions (Sections 2.6–2.9).
3. Multiparameter models, normal with unknown mean and variance, the multivariate normal distribution, multinomial models, election polling, bioassay. Computation and simulation from arbitrary posterior distributions in two parameters (Chapter 3).
4. Inference from large samples and comparison to standard non-Bayesian methods (Chapter 4).
5. Hierarchical models, estimating population parameters from data, rat tumor rates, SAT coaching experiments, meta-analysis (Chapter 5).
6. Model checking, posterior predictive checking, sensitivity analysis, model comparison and expansion, checking the analysis of the SAT coaching experiments (Chapter 6).
7. Data collection—ignorability, surveys, experiments, observational studies, unintentional missing data (Chapter 7).
8. General advice, connections to other statistical methods, examples of potential pitfalls of Bayesian inference (Chapters 8 and 9).
9. Computation: overview, uses of simulations, Gibbs sampling (Chapter 10, Sections 11.1–11.3).
10. Markov chain simulation (Sections 11.4–11.10, Appendix C).
11. Normal linear regression from a Bayesian perspective, incumbency advantage in Congressional elections (Chapter 14).
12. Hierarchical linear models, selection of explanatory variables, forecasting Presidential elections (Chapter 15).
13. Generalized linear models, police stops example, opinion polls example (Chapter 16).
14. Final weeks: topics from remaining chapters (including advanced computational methods, robust inference, mixture models, multivariate models, nonlinear models, missing data, and decision analysis).

Computer sites and contact details

Additional materials, including the data used in the examples, solutions to many of the end-of-chapter exercises, and any errors found after the book goes to press, are posted at `http://www.stat.columbia.edu/~gelman/`. Please send any comments to us at `gelman@stat.columbia.edu`, `sternh@uci.edu`, `jbcarlin@unimelb.edu.au`, or `rubin@stat.harvard.edu`.

Acknowledgments

We thank Stephen Ansolabehere, Adriano Azevedo, Jarrett Barber, Tom Belin, Suzette Blanchard, Brad Carlin, Alicia Carriquiry, Samantha Cook, Victor De Oliveira, David Draper, John Emerson, Steve Fienberg, Yuri Goegebeur, Daniel Gianola, David Hammill, Chuanpu Hu, Zaiying Huang, Yoon-Sook Jeon, Shane Jensen, Jay Kadane, Jouni Kerman, Gary King, Lucien Le Cam, Rod Little, Tom Little, Chuanhai Liu, Xuecheng Liu, Peter McCullagh, Mary Sara McPeek, Xiao-Li Meng, Baback Moghaddam, Olivier Nimeskern, Ali Rahimi, Thomas Richardson, Scott Schmidler, Andrea Siegel, Sandip Sinharay, Elizabeth Stuart, Andrew Swift, Francis Tuerlinckx, Iven Van Mechelen, Rob Weiss, Alan Zaslavsky, several reviewers, many other colleagues, and the students in Statistics 238, 242A, and 260 at Berkeley, Statistics 36-724 at Carnegie Mellon, Statistics 320 at Chicago, Statistics 220 at Harvard, Statistics 544 at Iowa State, and Statistics 6102 at Columbia, for helpful discussions, comments, and corrections. We especially thank Phillip Price and Radford Neal for their thorough readings of different parts of this book. John Boscardin deserves special thanks for implementing many of the computations for Sections 5.5, 6.8, 15.2, and 17.4. We also thank Chad Heilig for help in preparing tables, lists, and indexes. The National Science Foundation provided financial support through a postdoctoral fellowship and grants SBR-9223637, 9708424, DMS-9404305, 9457824, 9796129, and SES-9987748, 0084368. The computations and figures were done using the `S`, `S-Plus`, `R`, and `Bugs` computer packages (see Appendix C).

Many of the examples in this book have appeared elsewhere, in books and articles by ourselves and others, as we indicate in the bibliographic notes and exercises in the chapters where they appear. (In particular: Figures 1.3–1.5 are adapted from the *Journal of the American Statistical Association* **90** (1995), pp. 696, 702, and 703, and are reprinted with permission of the American Statistical Association. Figures 2.7 and 2.8 come from Gelman, A., and Nolan, D., *Teaching Statistics: A Bag of Tricks*, Oxford University Press (1992), pp. 14 and 15, and are reprinted with permission of Oxford University Press. Figures 20.8–20.10 come from the *Journal of the American Statistical Association* **91** (1996), pp. 1407 and 1409, and are reprinted with permission of the American Statistical Association. Table 20.1 comes from Berry, D., *Statistics: A Bayesian Perspective*, Duxbury Press (1996), p. 81, and is reprinted with permission of Brooks/Cole, a division of Thomson Learning. Figures 21.1 and 21.2 come from the *Journal of the American Statistical Association* **93** (1998) pp. 851 and 853, and are reprinted with permission of the American Statistical Association. Figures 22.1–22.3 are adapted from the *Journal of Business and Economic Statistics* **21** (2003), pp. 219 and 223, and are reprinted with permission of the American Statistical Association.)

Finally, we thank our spouses, Caroline, Nancy, Hara, and Kathryn, for their love and support during the writing and revision of this book.

Part I: Fundamentals of Bayesian Inference

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations. In Chapters 1–3, we introduce several useful families of models and illustrate their application in the analysis of relatively simple data structures. Some mathematics arises in the analytical manipulation of the probability distributions, notably in transformation and integration in multiparameter problems. We differ somewhat from other introductions to Bayesian inference by emphasizing stochastic simulation, and the combination of mathematical analysis and simulation, as general methods for summarizing distributions. Chapter 4 outlines the fundamental connections between Bayesian inference, other approaches to statistical inference, and the normal distribution. The early chapters focus on simple examples to develop the basic ideas of Bayesian inference; examples in which the Bayesian approach makes a practical difference relative to more traditional approaches begin to appear in Chapter 3. The major practical advantages of the Bayesian approach appear in hierarchical models, as discussed in Chapter 5 and thereafter.