CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Series Editors
C. Chatfield, *University of Bath, UK*
J. Zidek, *University of British Columbia, Canada*

The Analysis of Time Series —
An Introduction, Fifth Edition
C. Chatfield

Applied Bayesian Forecasting and Time
Series Analysis
A. Pole, M. West and J. Harrison

Applied Non-parametric Statistical
Methods, Second Edition,
P Sprent

Applied Statistics - Principles and
Examples
D.R. Cox and E.J. Snell

Bayesian Data Analysis
A. Gelman, J. Carlin, H. Stern and D. Rubin

Beyond ANOVA-Basics of Applied
Statistics
R.G. Miller, Jr.

Computer-Aided Multivariate Analysis,
Third Edition
A.A. Afifi and V.A. Clark

A Course in Categorical Data Analysis
T Leonard

A Course in Large Sample Theory
T .S. Ferguson

Data Driven Statistical Methods
P Sprent

Decision Analysis - A Bayesian Approach
J. 2 Smith

Elementary Applications of Probability
Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology-Study Design and
Data Analysis
M. Woodward

Essential Statistics, Third Edition
D.G. Rees

Interpreting Data -A First Course
in Statistics
A.J.B. Anderson

An Introduction to Generalized
Linear Models
A.J. Dobson

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods
and their Applications in Statistics
B.S. Everitt

Large Sample Methods in Statistics
R K. Sen and J. da Motta Singer

Markov Chain Monte Carlo - Stochastic
Simulation for Bayesian Inference
D. Gamerman

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic
Systems
V Kulkarni

Modelling Binary Data
D. Collett

Modelling Survival Data in Medical
Research
D. Collett

Multivariate Analysis of Variance and
Repeated Measures -A Practical
Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics-
A Practical Approach
B. Flury and H. Riedwyl

Practical Data Analysis for Designed
Experiments
B. S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Statistics for Medical Research
D.G. Altman

Probability-Methods and Measurement
A. O'Hagan

Problem Solving -A Statistician's Guide,
Second Edition
C. Chatfield

Randomization, Bootstrap and
Monte Carlo Methods in Biology,
Second Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber,
T.J. Sweeting and R.L. Smith

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and
Experimental Biology, Second Edition
R. Mead, R.N. Curnow and A.M. Hasted

Statistical Process Control-Theory and
Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics forAccountants, Fourth Edition
S. Letchford

Statistics for Technology —
A Course in Applied Statistics,
Third Edition
C. Chatfield

Statistics in Engineering
A Practical Approach
A.V. Metcalfe

Statistics in Research and Development,
Second Edition
R. Caulcutt

The Theory of Linear Models
B. Jorgensen

BAYES AND EMPIRICAL BAYES METHODS FOR DATA ANALYSIS

Second edition

Bradley P. Carlin

Professor, Division of Biostatistics,
School of Public Health, University of Minnesota,
Minneapolis, MN

and

Thomas A. Louis

Senior Statistician,
The RAND Corporation,
Santa Monica, CA

CHAPMAN & HALL/CRC

A CRC Press Company

Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Carlin, Bradley P.

Bayes and empirical bayes methods for data analysis / Bradley P. Carlin and Thomas A. Louis--2nd ed.

p. cm. (Texts in statistical science series)

Previous ed.: New York: Chapman & Hall, 1996.

Includes bibliographical references and indexes.

ISBN 58488-170-4

1. Bayesian statistical decision theory. I. Louis, Thomas A., II. Title. III. Texts in statistical science

QA279.5.C36 2000

519'.5842-dc21 00-038242

CIP

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Visit the CRC Press Web site at www.crcpress.com

International Standard Book Number 1-58488-170-4

Library of Congress Card Number 00-38242

Printed in the United States of America 3 4 5 6 7 8 9 0

Printed on acid-free paper

Contents

TO NATHAN, JOSHUA, SAMUEL, ERICA, AND MARGIT

Preface to the Second Edition

Interest in and the resulting impact of Bayes and empirical Bayes (EB) methods continues to grow. An impressive expansion in the number of Bayesian journal articles, conference presentations, courses, seminars, and software packages has occurred in the four years since our first edition went to press. Perhaps more importantly. Bayesian methods now find application in a large and expanding number of areas where just a short time ago their implementation and routine use would have seemed a laudable yet absurdly unrealistic goal. The allocation of billions of U.S. federal dollars now depends on Bayesian estimates of population characteristics (e.g. poverty) in small geographic areas. The FDA now not only permits but encourages Bayesian designs and analyses in applications seeking approval for new medical devices such as pacemakers, an area where large clinical trials are infeasible but directly relevant historical information is often available. The Microsoft Windows Office Assistant is based on a Bayesian artificial intelligence algorithm; similarly. a prominent movie-rental chain now offers a web-based Bayesian expert system to help you choose which film to rent, based on your general preferences and the names of other films you have enjoyed. No longer merely of academic interest, Bayes and EB methods have assumed their place in the toolkit of the well-equipped applied statistician.

A principal reason for the ongoing expansion in the Bayes and EB statistical presence is of course the corresponding expansion in readily-available computing power, and the simultaneous development in Markov chain Monte Carlo (MCMC) methods and software for harnessing it. Our desire to incorporate many recent developments in this area provided one main impetus for this second edition. For example, we now include discussions of reversible jump MCMC, slice sampling, structured MCMC, and other new computing methods and software packages. Feedback from readers and practitioners also helped us identify important application areas given little or no attention in the first edition. As such, we have expanded our coverage of Bayesian methods in spatial statistics, sequential analysis, model choice, and experimental design, especially sample size estimation for clinical trials. This edition also features an updated reference list, expanded and improved subject and author indices, and corrections of many (though surely not all) typographical errors.

Still, perhaps our main motivation in preparing this second edition was to improve the book's effectiveness as a course text. The most significant change in this regard is the moving of the decision-theoretic material from Chapter 1 to a location (new Appendix B) specifically devoted to this topic. Chapter 1 now includes several vignettes and elementary examples which illustrate Bayesian thinking and benefits. Our intent is to provide a gentler and more direct introduction to the "nuts and bolts" issues of Bayesian analysis in Chapter 2.

Besides these structural changes, we have attempted to enhance the book's teaching value in other ways. These include a gentler, better introduction to Gibbs sampling in Subsection 5.4.2, an explicit description of how to estimate MCMC standard errors in Subsection 5.4.5, a greatly revised treatment of Bayesian model choice in Sections 6.3, 6.4, and 6.5, actual BUGS and WinBUGS code to do several of the examples in the book, and the addition of several new homework problems and solutions.

As with the first edition, this revision presupposes no previous exposure to Bayes or EB methods, but only a solid, Master's-level understanding of mathematical statistics - say, at the level of Hogg and Craig (1978), Mood, Graybill, and Boes (1974), or Casella and Berger (1990). A course on the basics of modern applied Bayesian methods might cover only Chapters 1, 2, 5, and 6, with supplemental material from Chapters 3 (EB), 7 (special topics), and 8 (case studies) as time permitted. A more advanced course might instead refer to Appendix B after Chapter 2, then return to the EB and performance material in Chapters 3 and 4, since this material is directly motivated by decision theory. The advanced computing, methods, and data analysis material in Chapters 5 and beyond then follows naturally. In our course in the Division of Biostatistics at the University of Minnesota we cover the entire book in a four-credit-hour, single-semester (15 week) course; the aforementioned "short course" could surely be completed in far less time. See `http://www.biostat.umn.edu/~brad/` on the web for many of our datasets and other teaching-related information.

We owe a debt of gratitude to those who helped in our revision process. Sid Chib, Mark Glickman, Betz Halloran, Siem Heisterkamp, David Higdon, Jim Hodges, Giovanni Parmigiani, Jeff Rosenthal, David Spiegelhalter, Melanie Wall, and Larry Wasserman all provided valuable feedback, spotting errors and suggesting directions our revision might take. Cong Han and Li Zhu provided invaluable technical support. We are especially grateful to Antonietta Mira for explaining slice sampling to us. Finally, we thank our families, whose ongoing love and support made all of this possible.

BRADLEY P. CARLIN
THOMAS A. LOUIS

Minneapolis, Minnesota
April 2000

Preface to the First Edition

Bayes and empirical Bayes methods have been shown to have attractive properties for both the Bayesian and the frequentist. However, a combination of philosophical posturing and difficulties in implementation have worked against these approaches becoming standard data-analytic tools. Our intent in this book is to help redress this situation by laying out the formal basis for Bayes and empirical Bayes (EB) approaches, evaluating their Bayes and frequentist properties, and showing how they can be implemented in complicated model settings. Our hope is that readers will better understand the benefits (and limitations) of Bayes and EB, and will increasingly use these approaches in their own work.

Though there are several books on the philosophical and mathematical aspects of Bayes and EB methods. there are none we know of that strongly link modern developments to applications while also showing the several practical benefits of these approaches irrespective of one's philosophical viewpoint. We believe that progress in computing and the attendant modeling flexibility sets the stage for Bayes and empirical Bayes methods to be of substantial benefit for the applied statistician.

We intend that this book will help realize this benefit by avoiding excessive philosophical nitpicking, and instead show how Bayes and EB procedures operate in both theory and practice. The book will satisfy the growing need for a ready reference to Bayes and EB that can be read and appreciated by the practicing statistician as well as the second-year graduate student, and a text that casts the formal development in terms of applied linkages. The book is intended for those with a good Master's-level understanding of mathematical statistics, but we assume no formal exposure to Bayes or empirical Bayes methods. It should be appropriate for use both as a text and as a reference. Our somewhat informal development is designed to enhance understanding; mathematical proofs are kept to a minimum. We hope that the mathematical level and breadth of examples will recruit students and teachers not only from statistics and biostatistics, but from a broad range of fields, including for example epidemiology, econometrics, education, psychology, and sociology.

Bayes and EB touches much of modern statistical practice, and we cannot include all relevant theory and examples. For example, the theory of diagnostic tests is best presented in a Bayesian fashion, and Bayesian formalism lies behind all random effects and errors-in-variables models. We have dealt with this breadth by including models and examples that effectively communicate the approach, but make no claim to being comprehensive. For example, this book does little more than mention nonparametric Bayesian approaches, such as the use of Dirichlet process priors. even though they are a subject of greatly increased interest of late thanks to the addition of Markov chain Monte Carlo (MCMC) methods to the Bayesian toolkit.

Philosophically, two principal themes dominate the book: that effective statistical procedures strike a balance between variance and bias. and that Bayesian formalism in a properly robust form produces procedures that achieve this balance. This success applies whether one evaluates methods as a Bayesian (considering parameters as random variables and the data as fixed), or as a frequentist (over a range of fixed parameter values). With this in mind, we begin in Chapter 1 by outlining the decision-theoretic tools needed to compare procedures, though unlike more traditional Bayesian texts (e.g., DeGroot, 1970; Berger, 1985; Robert, 1994) we do not dwell on them. After presenting the basics of the Bayes and EB approaches in Chapters 2 and 3, respectively, Chapter 4 evaluates the frequentist and empirical Bayes (preposterior) performance of these approaches in a variety of settings. Of course, no single approach can be universally best. and we identify both virtues and drawbacks. This sharpens our focus on generating procedures that do not rely heavily on prior information (i.e., that enjoy good frequentist and EB properties over a broad range of parameter values).

We intend for the book to have a very practical focus, offering real solution methods to researchers with challenging problems. This is especially true of the book's second half, which begins with an extensive discussion of modern Bayesian computational methods in Chapter 5. These methods figure prominently in modern tools for such data analytic tasks as model criticism and selection, which are described in Chapter 6. Guidance on the Bayes/EB implementation of a generous collection of special methods and models is given in Chapter 7. Finally, we close in Chapter 8 with three fully worked case studies of real datasets. These studies incorporate tools from a variety of statistical subfields, including longitudinal analysis, survival models, sequential analysis, and spatio-temporal mapping.

The book concludes with three useful appendices. Appendix A contains a brief summary of the distributions used in the book, highlighting the typical Bayes/EB role of each. For those wishing to avoid as much computer programming as possible, Appendix B provides a guide to the commercial and noncommercial software available for performing Bayesian analyses, indicating the level of model complexity each is capable of handling. While guides such as ours become outdated with remarkable rapidity, we believe it to be fairly complete as of the time of our writing, and hope it will at least serve as a starting point for the beginning applied Bayesian. Finally, Appendix C contains solutions to several of the exercises in each of the book's chapters.

We encourage the student to actually try the exercises at the end of each chapter, since only through practice will the student gain a true appreciation for the (often computationally intensive) methods we advocate. We caution that each chapter's exercises are not necessarily arranged in order of increasing difficulty, but rather in roughly the same order as the topics in that chapter, oftentimes with the more theoretical and methodological questions preceding the more computational and data-analytic ones.

We are very grateful to a variety of persons who helped make this book possible. Deb Sampson typed a significant portion of the first draft, and also proved invaluable in her role as "LATEX goddess." Wei Shen, Hong Xia, and Dandan Zheng spent hours writing computer simulation code and drawing many of the *S+* plots seen throughout the book. The entire 1995 and 1996 winter quarter "Bayes and Empirical Bayes Methods" classes served as aggresive copy-editors, finding flawed homework problems, missing references, and an embarrassing number of misconstructed sentences. Finally, we thank Caroline S. Carlin. F.S.A. and Karen Seashore Louis, Ph.D. for their unending patience and good humor. So scarred are they from the experience that both now use words like "posterior sensitivity" in polite conversation.

BRADLEY P. CARLIN — Minneapolis, Minnesota
THOMAS A. LOUIS — March 1996

CHAPTER 1

Approaches for statistical inference

1.1 Introduction

The practicing statistician faces a variety of challenges: designing complex studies, summarizing complex data sets, fitting probability models, drawing conclusions about the present, and making predictions for the future. Statistical studies play an important role in scientific discovery, in policy formulation, and in business decisions. Applications of statistics are ubiquitous, and include clinical decision making, conducting an environmental risk assessment, setting insurance rates, deciding whether (and how) to market a new product, and allocating federal funds. Currently, most statistical analyses are performed with the help of commercial software packages, most of which use methods based on a classical, or *frequentist,* statistical philosophy. In this framework, maximum likelihood estimates (MLEs) and hypothesis tests based on p-values figure prominently.

Against this background, the *Bayesian* approach to statistical design and analysis is emerging as an increasingly effective and practical alternative to the frequentist one. Indeed, due to computing advances which enable relevant Bayesian designs and analyses, the philosophical battles between frequentists and Bayesians that were once common at professional statistical meetings are being replaced by a single, more eclectic approach. The title of our book makes clear which philosophy and approach we prefer; that you are reading it suggests at least some favorable disposition (or at the very least, curiosity) on your part as well. Rather than launch headlong into another spirited promotional campaign for Bayesian methods, we begin with motivating vignettes, then provide a basic introduction to the Bayesian formalism followed by a brief historical account of Bayes and frequentist approaches, with attention to some controversies. These lead directly to our methodological and applied investigations.

1.2 Motivating vignettes

1.2.1 Personal probability

Suppose you have submitted your first manuscript to a journal and have assessed the chances of its being accepted for publication. This assessment uses information on the journal's acceptance rate for manuscripts like yours (let's say around 30%), and your evaluation of the manuscript's quality. Subsequently, you are informed that the manuscript has been accepted (congratulations!). What is your updated assessment of the probability that your *next* submission (on a similar topic) will be accepted?

The direct estimate is of course 100% (thus far, you have had one success in one attempt), but this estimate seems naive given what we know about the journal's overall acceptance rate (our external, or *prior,* information in this setting). You might thus pick a number smaller than 100%; if so, you are behaving as a Bayesian would because you are adjusting the (unbiased, but weak) direct estimate in the light of your prior information. This ability to formally incorporate prior information into an analysis is a hallmark of Bayesian methods, and one that frees the analyst from ad hoc adjustments of results that "don't look right."

1.2.2 Missing data

Consider Table 1.1, reporting an array of stable event prevalence or incidence estimates scaled per 10,000 population, with one value (indicated by "*") missing at random. The reader may think of them as geographically aligned disease prevalences, or perhaps as death rates cross-tabulated by clinic and age group.

79	87	83	80	78
90	89	92	99	95
96	100	*	110	115
101	109	105	108	112
96	104	92	101	96

Table 1.1 *An array of well-estimated rates per 10,000 with one estimate missing*

With no direct information for *, what would you use for an estimate? Does 200 seem reasonable? Probably not, since the unknown rate is surrounded by estimates near 100. To produce an estimate for the missing cell you might fit an additive model (rows and columns) and then use the model to impute a value for *, or merely average the values in surrounding cells. These are two examples of *borrowing information.* Whatever your approach, some number around 100 seems reasonable.

Now assume that we obtain data for the * cell and the estimate is, in fact, 200, based on 2 events in a population of 100 (200 = 10000 x 2/100). Would you now estimate * by 200 (a very unstable estimate based on very little information), when with no information a moment ago you used 100? While 200 is a perfectly valid estimate (though its uncertainty should be reported), some sort of *weighted average* of this direct estimate (200) and the indirect estimate you used when there was no direct information (100) seems intuitively more appealing. The Bayesian formalism allows just this sort of natural compromise estimate to emerge.

Finally, repeat this mental exercise assuming that the direct estimate is still 200 per 10,000, but now based on 20 events in a population of 1000, and then on 2000 events in a population of 100,000. What estimate would you use in each case? Bayes and empirical Bayes methods structure this type of statistical decision problem, automatically giving increasing weight to the direct estimate as it becomes more reliable.

1.2.3 Bioassay

Consider a carcinogen bioassay where you are comparing a control group (C) and an exposed group (E) with 50 rodents in each (see Table 1.2). In the control group, 0 tumors are found; in the exposed group, there are 3, producing a non-significant, one-sided Fisher exact test p-value of approximately 0.125. However, your colleague, who is a veterinary pathologist, states, "I don't know about statistical significance, but three tumors in 50 rodents is certainly *biologically* significant!"

	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
	50	50	100

Table 1.2 *Hypothetical bioassay results; one-sided p* = 0.125

This belief may be based on information from other experiments in the same lab in the previous year in which the tumor has never shown up in control rodents. For example, if there were 400 historical controls in addition to the 50 concurrent controls, none with a tumor, the one sided p-value becomes 0.001 (see Table 1.3). Statistical and biological significance are now compatible. Still, it is generally inappropriate simply to pool historical and concurrent information. However, Bayes and empirical Bayes methods may be used to structure a valid synthesis; see, for example, Tarone (1982), Dempster et al. (1983), Tamura and Young (1986), Louis and Bailey (1990), and Chen et al. (1999).

	C	E	Total
Tumor	0	3	3
No Tumor	450	47	497
	450	50	500

Table 1.3 *Hypothetical bioassay results augmented by 400 historical controls, none with a tumor, one-sided* $p = 0.001$.

1.2.4 Attenuation adjustment

In a standard errors-in-variables simple linear regression model, the least squares estimate of the regression slope (β) is biased toward 0, an example of *attenuation*. More formally suppose the true regression is $Y = \beta x + \epsilon$, $\epsilon \sim N(0, \sigma_\epsilon^2)$ but Y is regressed not on x but on $X \equiv x + \delta$, where $\delta \sim N(0, \sigma_\delta^2)$. Then the least squares estimate has expectation $E[\hat{\beta}] \approx \rho\beta$, with $\rho = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\delta^2} \leq 1$. If ρ is known or well-estimated, one can correct for attenuation and produce an unbiased estimate by using $\hat{\beta}/\rho$ to estimate β.

Though unbiasedness is an attractive property, especially when the standard error associated with the estimate is small, in general it is less important than having the estimate "close" to the true value. The expected squared deviation between the true value and the estimate (*mean squared error,* or MSE) provides an effective measure of proximity. Fortunately for our intuition, MSE can be written as the sum of an estimator's sampling variance and its squared bias,

$$\text{MSE} = \text{variance} + (\text{bias})^2 . \qquad (1.1)$$

The unbiased estimate sets the second term to 0, but it can have a verv large MSE relative to other estimators; in this case, because dividing $\hat{\beta}$ by ρ inflates the variance as the price of eliminating bias.

In the exercises for Appendix B, we ask that you construct a Bayesian estimator that strikes an effective tradeoff between variance and bias. It goes part way toward complete elimination of bias, with the degree of adjustment depending on the relative size of the variance and the bias.

1.3 Defining the approaches

Three principal approaches to inference guide modern data analysis: frequentist, Bayesian, and likelihood. We now briefly describe each in turn.

The *frequentist* evaluates procedures based on imagining repeated sampling from a particular model (the *likelihood*) which defines the probability distribution of the observed data conditional on unknown parameters. Properties of the procedure are evaluated in this repeated sampling framework for fixed values of unknown parameters; good procedures perform well over a broad range of parameter values.

The *Bayesian* requires a sampling model and, in addition, a *prior* distribution on all unknown quantities in the model (parameters and missing data). The prior and likelihood are used to compute the conditional distribution of the unknowns given the observed data (the *posterior* distribution), from which all statistical inferences arise. Allowing the observed data to play some role in determining the prior distribution produces the *empirical Bayes* (EB) approach. The Bayesian evaluates procedures for a repeated sampling experiment of unknowns drawn from the posterior distribution for a given data set. The empirical Bayesian may also evaluate procedures under repeated sampling of *both* the data and the unknowns from their joint distribution.

Finally, the "*likelihoodist*" (or "*Fisherian*")develops a sampling model but not a prior, as does the frequentist. However, inferences are restricted to procedures that use the data only as reported by the likelihood, as a Bayesian would. Procedure evaluations can be from a frequentist, Bayesian, or EB point of view.

As presented in Appendix B, the frequentist and Bayesian approaches can be connected in a *decision-theoretic* framework, wherein one considers a model where the unknown parameters and observed data have a joint distribution. The frequentist conditions on parameters and replicates (integrates) over the data; the Bayesian conditions on the data and replicates (integrates) over the parameters. *Preposterior* evaluations integrate over both parameters and data, and can be represented as frequentist performance averaged over the prior distribution, or as Bayesian posterior performance averaged over the marginal sampling distribution of the observed data (the sampling distribution conditional on parameters, averaged over the prior). Preposterior properties are relevant to both Bayesians and frequentists (see Rubin, 1984).

Historically, frequentists have criticized Bayesian procedures for their inability to deal with all but the most basic examples, for overreliance on computationally convenient priors, and for being too fragile in their dependence on a specific prior (i.e., for a lack of *robustness* in settings where the data and prior conflict). Bayesians have criticized frequentists for failure to incorporate relevant prior information, for inefficiency, and for *incoherence* (i.e., a failure to process available information systematically, as a Bayesian approach would). Recent computing advances have all but eliminated constraints on priors and models, but leave open the more fundamental critique of non-robustness. In this book, therefore, we shall often seek the middle ground: Bayes and EB procedures that offer many of the Bayes advantages, but do so without giving up too much frequentist robustness. Procedures occupying this middle ground can be thought of as having good performance over a broad range of prior distributions, but not so broad as to include all priors (which would in turn require good performance over all possible parameter values).

1.4 The Bayes-frequentist controversy

While probability has been the subject of study for hundreds of years (most notably by mathematicians retained by rich noblemen to advise them on how to maximize their winnings in games of chance), statistics is a relatively young field. Linear regression first appeared in the work of Francis Galton in the late 1800s, with Karl Pearson adding correlation and goodness-of-fit measures around the turn of the century. The field did not really blossom until the 1920s and 1930s, when R.A. Fisher developed the notion of likelihood for general estimation, and Jerzy Neyman and Egon Pearson developed the basis for classical hypothesis testing. A flurry of research activity was energized by the Second World War, which generated a wide variety of difficult applied problems and the first substantive government funding for their solution in the United States and Great Britain.

By contrast, Bayesian methods are much older, dating to the original 1763 paper by the Rev. Thomas Bayes, a minister and amateur mathematician. The area generated some interest by Laplace, Gauss, and others in the 19th century, but the Bayesian approach was ignored (or actively opposed) by the statisticians of the early 20th century. Fortunately, during this period several prominent non-statisticians, most notably Harold Jeffreys (a physicist) and Arthur Bowley (an econometrician), continued to lobby on behalf of Bayesian ideas (which they referred to as "inverse probability"). Then, beginning around 1950, statisticians such as L.J. Savage, Bruno de Finetti, Dennis Lindley, Jack Kiefer, and many others began advocating Bayesian methods as remedies for certain deficiencies in the classical approach. The following example illustrates such a deficiency in interval estimation.

Example 1.1 Suppose $X_i \stackrel{iid}{\sim} N(\theta, \sigma^2)$, $i = 1, \ldots, n$, where N denotes the normal (Gaussian) distribution and iid stands for "independent and identically distributed." We desire a 95% interval estimate for the population mean θ. Provided n *is* sufficiently large (say, bigger than 30), a classical approach would use the confidence interval

$$\delta(\mathbf{x}) = \bar{x} \pm 1.96 s/\sqrt{n} \, ,$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, $\bar{x}$ is the sample mean, and s is the sample standard deviation.

This interval has the property that. on the average over repeated application, $\delta(\mathbf{x})$ will fail to capture the true mean θ only 5% of the time. An alternative interpretation is that, *before* any data are collected, the probability that the interval contains the true value is 0.95. This property is attractive in the sense that it holds for *all* true values of θ and σ^2.

On the other hand, its use in any single data-analytic setting is somewhat difficult to explain and understand. After collecting the data and computing $\delta(\mathbf{x})$ the interval either contains the true θ or it does not; its coverage probability is not 0.95, but either 0 or 1. After observing **x**, a statement like, "the true θ has a 95% chance of falling in $\delta(\mathbf{x})$," is not valid, though most people (including most statisticians irrespective of their philosophical approach) interpret a confidence interval in this way. Thus for the frequentist, "95%" is not a conditional coverage probability, but rather a tag associated with the interval to indicate either how it is likely to perform before we evaluate it, or how it would perform over the long haul. A 99% frequentist interval would be wider, a 90% interval narrower, but, conditional on **x**, all would have coverage probability 0 or 1. ■

By contrast, Bayesian confidence intervals (known as "credible sets," and discussed further in Subsection 2.3.2) are free of this awkward frequentist interpretation. For example, conditional on the observed data, the probability is 0.95 that θ is in the 95% credible interval. Of course, this natural interpretation comes at the price of needing to specify a prior distribution for θ so we must provide additional structure to obtain ease of interpretation.

The Neyman-Pearson testing structure can also lead to some very odd results, which have been even more heavily criticized by Bayesians.

Example 1.2 Consider the following simple experiment, originally suggested by Lindley and Phillips (1976), and reprinted many times. Suppose in 12 independent tosses of a coin, I observe 9 heads and 3 tails. I wish to test the null hypothesis $H_0 : \theta = 1/2$ versus the alternative hypothesis $H_a : \theta > 1/2$, where θ is the true probability of heads. Given only this much information, two choices for the sampling distribution emerge as candidates:

1. *Binomial:* The number $n = 12$ tosses was fixed beforehand, and the random quantity X was the number of heads observed in the n tosses. Then $X \sim Bin(12, \theta)$ and the likelihood function is given by
$$L_1(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3 \; . \qquad (1.2)$$

2. *Negative binomial:* Data collection involved flipping the coin until the third tail appeared. Here, the random quantity X is the number of heads required to complete the experiment, so that $X \sim NegBin(r = 3, \theta)$, with likelihood function given by
$$L_2(\theta) = \binom{r+x-1}{x}\theta^x(1-\theta)^r = \binom{11}{9}\theta^9(1-\theta)^3 \; . \qquad (1.3)$$

Under either of these two alternatives, we can compute the p-value corresponding to the rejection region, "Reject H_0 if $X \geq c$." Doing so using the binomial likelihood (1.2) we obtain

$$\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = .075 \, ,$$

while for the negative binomial likelihood (1.3),

$$\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{\infty} \binom{2+j}{j} \theta^j (1-\theta)^3 = .0325 \, .$$

Thus, using the "usual" Type I error level $\alpha = .05$, we see that the two model assumptions lead to two different decisions: we would reject. H_0 if X were assumed negative binomial, but not if it were assumed binomial. But there is no information given in the problem setting to help us make this determination, so it is not clear which analysis the frequentist should regard as "correct." In any case, assuming we trust the statistical model, it does not seem reasonable that how the experiment was monitored should have any bearing on our decision; surely only its *results* are relevant! Indeed, the likelihood functions tell a consistent story, since (1.2) and (1.3) differ only by a multiplicative constant that does not depend on θ. ■

A Bayesian explanation of what went wrong in the previous example would be that the Neyman-Pearson approach allows *unobserved* outcomes to affect the rejection decision. That is, the probability of X values "more extreme" than 9 (the value actually observed) was used as evidence against H_0 in each case, even though these values did not occur. More formally, this is a violation of a statistical axiom known as the *Likelihood Principle,* a notion present in the work of Fisher and Barnard, but not precisely- defined until the landmark paper by Birnbaum (1962). In a nutshell. the Likelihood Principle states that once the data value x has been observed, the likelihood function $L(\theta|x)$ contains all relevant experimental information delivered by x about the unknown parameter θ. In the previous example, L_1 and L_2 are proportional to each other as functions of θ, hence are equivalent in terms of experimental information (recall that multiplying a likelihood function by an arbitrary function $h(x)$ does not change the MLE B). Yet in the Neyman-Pearson formulation, these equivalent likelihoods lead to two different inferences regarding θ. Put another way, frequentist test results actually depend not only on what x was observed, but on how the experiment was stopped.

Some statisticians attempt to defend the results of Example 1.2 by arguing that all aspects of the design of an experiment *are* relevant pieces of information even after the data have been collected, or perhaps that the Likelihood Principle is itself flawed (or at least should not be considered sacrosanct). We do not delve deeper into this foundational discussion, but refer the interested reader to the excellent monograph by Berger and

Wolpert (1984) for a presentation of the consequences, criticisms, and defenses of the Likelihood Principle. Violation of the Likelihood Principle is but one of the possibly anomalous properties of classical testing methods; more are outlined later in Subsection 2.3.3, where it is also shown that Bayesian hypothesis testing methods overcome these difficulties.

If Bayesian methods offer a solution to these and other drawbacks of the frequentist approach that were publicized several decades ago, it is perhaps surprising that Bayesian methodology did not make bigger inroads into actual statistical practice until only recently. There are several reasons for this. First, the initial staunch advocates of Bayesian methods were primarily subjectivists, in that they argued forcefully that all statistical calculations should be done only after one's own personal prior beliefs on the subject had been carefully evaluated and quantified. But this raised concerns on the part of classical statisticians (and some Bayesians) that the results obtained would not be objectively valid, and could be manipulated in any way the statististician saw fit. (The subjectivists' reply that frequentist methods were invalid anyway and should be discarded did little to assuage these concerns.) Second, and perhaps more important from an applied standpoint, the Bayesian alternative, while theoretically simple, required evaluation of complex integrals in even fairly rudimentary problems. Without inexpensive, high-speed computing, this practical impediment combined with the theoretical concerns to limit the growth of realistic Bayesian data analysis. However, this growth finally did occur in the 1980s, thanks to a more objective group of Bayesians with access to inexpensive, fast computing. The objectivity issue is discussed further in Chapter 2, while computing is the subject of Chapter 5.

These two main concerns regarding routine use of Bayesian analysis (i.e., that its use is not easy or automatic, and that it is not always clear how objective results may be obtained) were raised in the widely-read paper by Efron (1986). While the former issue has been largely resolved in the intervening years thanks to computing advances, the latter remains a challenge and the subject of intense ongoing research (see Subsection 2.2.3 for an introduction to this area). Still, we contend that the advantages in using Bayes and EB methods justify the increased effort in computation and prior determination. Many of these advantages are presented in detail in the popular textbook by Berger (1985, Section 4.1); we quickly review several of them here.

1. Bayesian methods provide the user with the ability to formally incorporate prior information.

2. Inferences are conditional on the *actual* data.

3. The reason for stopping the experimentation does not affect Bayesian inference (as shown in Example 1.2).

4. Bayesian answers are more easily interpretable by nonspecialists (as shown in Example 1.1).
5. All Bayesian analyses follow *directly* from the posterior distribution; no separate theories of estimation, testing, multiple comparisons, etc. are needed.
6. Any question can be *directly* answered through Bayesian analysis.
7. Bayes and EB procedures automatically possess numerous optimality properties.

As an example of the sixth point, to investigate the bioequivalence of two drugs, we would need to test $H_0 : \theta_1 \neq \theta_2$ versus $H_a : \theta_1 = \theta_2$. That is. we must reverse the traditional null and alternative roles. since the hypothesis we hope to reject is that the drugs are *different,* not that they are the same. This reversal turns out to make things quite awkward for traditional testing methods (see e.g., Berger and Hsu, 1996 for a careful review), but not for Bayesian testing methods, since they treat the null and alternative hypotheses equivalently; they really can "accept" the null. rather than merely "fail to reject" it. Finally, regarding the seventh point. Bayes procedures are typically consistent, can automatically impose parsimony in model choice, and can even define the class of optimal *frequentist* procedures, thus "beating the frequentist at his own game." We return to this final issue (frequentist motivations for using Bayes and EB procedures) in Chapter 4.

1.5 Some basic Bayesian models

The most basic Bayesian model has two stages, with a likelihood specification $Y|\theta \sim f(y|\theta)$ and a prior specification $\theta \sim \pi(\theta)$, where either Y or θ can be vectors. In the simplest Bayesian analysis, π is assumed known, so that by probability calculus, the posterior distribution of θ is given by

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{m(y)} , \tag{1.4}$$

where

$$m(y) = \int f(y \mid \theta)\pi(\theta)d\theta , \tag{1.5}$$

the *marginal* density of the data y. Equation (1.4) is a special case of *Bayes' Theorem,* the general form of which we present in Section 2.1. For all but rather special choices of f and π (see Subsection 2.2.2), evaluating integrals such as (1.5) used to be difficult or impossible, forcing Bayesians into unappealing approximations. However, recent developments in Monte Carlo computing methods (see Chapter 5) allow accurate estimation of such integrals, and have thus enabled advanced Bayesian data analysis.

1.5.1 A Gaussian/Gaussian (normal/normal) model

We now consider the case where both the prior and the likelihood are Gaussian (normal) distributions,* namely,

$$\begin{aligned} \theta &\sim N(\mu, \tau^2) \qquad (1.6) \\ Y \mid \theta &\sim N(\theta, \sigma^2). \end{aligned}$$

The marginal distribution of Y given in (1.5) turns out to be $N(\mu, \sigma^2+\tau^2)$, and the posterior distribution (1.4) is also Gaussian with mean and variance

$$E(\theta \mid Y) = B\mu + (1-B)Y \qquad (1.7)$$

$$Var(\theta \mid Y) = (1-B)\sigma^2 \qquad (1.8)$$

where $B = \sigma^2/(\sigma^2+\tau^2)$. Since $0 \leq B \leq 1$, the posterior mean is a weighted average of the prior mean μ and the direct estimate Y; the direct estimate is pulled back (or *shrunk)* toward the prior mean. Moreover, the weight on the prior mean B depends on the relative variability of the prior distribution and the likelihood. If σ^2 is large relative to τ^2 (i.e., our prior knowledge is more precise than the data information), then B is close to 1, producing substantial shrinkage. If σ^2 is small (i.e., our prior knowledge is imprecise relative to the data information), B is close to 0 and the direct estimate is moved very little toward the prior mean. As we show in Chapter 4, this shrinkage provides an effective tradeoff between variance and bias, with beneficial effects on the resulting mean squared error; see equation (1.1).

If one is willing to assume that the structure (1.6) holds, but that the prior mean μ and variance τ^2 are unknown, then *empirical* Bayes or *hyper-prior* Bayes methods can be used; these methods are the subject of much of the remainder of this book.

1.5.2 A beta/binomial model

Next, consider applying the Bayesian approach given in (1.4)-(1.5) to estimating a binomial success probability. With Y the number of events in n independent trials and θ the event probability, the sampling distribution is

$$P(Y=y \mid \theta) = f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} .$$

In order to obtain a closed form for the marginal distribution, we use the *Beta(a, b)* prior distribution for θ (see Section A.2 in Appendix A). For convenience we reparametrize from (a, b) to (μ, M) where $\mu = a/(a+b)$, the prior mean, and $M = a + b$, a measure of prior precision. More specifically, the prior variance is then $\mu(1-\mu)/(M+1)$, a decreasing function of M. As shown in detail in Subsection 3.3.2, the marginal distribution of Y is

* Throughout the book, we use the terms "normal" and "Gaussian" interchangeably.

then beta-binomial with $E\,(Y/n) = \mu$ and

$$Var\left(\frac{Y}{n}\right) = \frac{\mu(1-\mu)}{n}\left[1 + \frac{n-1}{M+1}\right] .$$

The term in square brackets is known variously as the "variance inflation factor," "design effect," or "component of extra-binomial variation."

The posterior distribution of θ given Y in (1.4) is again *Beta* with mean

$$\begin{aligned} \hat{\theta} &= \frac{M}{M+n}\mu + \frac{n}{M+n}\left(\frac{Y}{n}\right), \qquad (1.9) \\ &= \mu + \frac{n}{M+n}\left[\left(\frac{Y}{n}\right) - \mu\right] \end{aligned}$$

and variance $Var(\theta \mid Y) = [\hat{\theta}(1-\hat{\theta})]/(M+n)$. Note that. similar to the posterior mean (1.7) in the Gaussian/Gaussian example, $\hat{\theta}$ is a weighted average of the prior mean μ and the maximum likelihood estimate Y/n, with weight depending on the relative size of M (the information in the prior) and n (the information in the data).

Discussion

Statistical decision rules can be generated by any philosophy under any collection of assumptions. They can then be evaluated by any criteria. even those arising from an utterly different philosophy. We contend (and the subsequent chapters will show) that the Bayesian approach is an excellent "procedure generator," even if one's evaluation criteria are frequentist provided that the prior distributions introduce only a small amount of information. This agnostic view considers features of the prior (possibly the entire prior) as "tuning parameters" that can be used to produce a decision rule with broad validity. The Bayesian formalism will be even more effective if one desires to structure an analysis using either personal opinion or objective information external to the current data set. The Bayesian approach also encourages documenting assumptions and quantifying uncertainty. Of course, no approach automatically produces broadly valid inferences, even in the context of the Bayesian models. A procedure generated by a high-information prior with most of its mass far from the truth will perform poorly under both Bayesian and frequentist evaluations.

The experienced statistician prefers design and analysis methods that strike an effective tradeoff between efficiency and robustness. Striking a balance is of principal importance, irrespective of the underlying philosophy. For example, in estimation central focus should be on reduction of MSE and related performance measures through a tradeoff between variance and bias. This concept is appropriate for both frequentists and Bayesians. In this context our strategy is to use the Bayesian formalism to reduce MSE even when evaluations are frequentist. In a broad range of situations, one can gain much of the Bayes or empirical Bayes advantage while retaining frequentist robustness.

Importantly, the Bayesian formalism properly propagates uncertainty through the analysis enabling a more realistic (typically inflated) assessment of the variability in estimated quantities of interest. Also, the formalism structures the analysis of complicated models where intuition may produce faulty or inefficient approaches. This structuring becomes especially important in multiparameter models in which the Bayesian approach requires that the joint posterior distribution of all parameters structure all analyses. Appropriate marginal distributions focus on specific parameters, and this integration ensures that all uncertainties influence the spread and shape of the marginal posterior.

1.6 Exercises

1. Let θ be the true proportion of men over the age of 40 in your community with hypertension. Consider the following "thought experiment"

   (a) Though you may have little or no expertise in this area, give an initial point estimate of θ .

   (b) Now suppose a survey to estimate 0 is established in your community, and of the first 5 randomly selected men, 4 are hypertensive. How does this information affect. your initial estimate of θ ?

   (c) Finally, suppose that at the survey's completion, 400 of 1000 men have emerged as hypertensive. Now what is your estimate of θ ?

   What guidelines for statistical inferential systems are suggested by your answers?

2. Repeat the journal publication thought problem from Subsection 1.2.1 for the situation where

   (a) you have won a lottery on your first try.

   (b) you have correctly predicted the winner of the first game of the World Series (professional baseball).

3. Assume you have developed predictive distributions of the length of time it takes to drive to work, one distribution for Route A and one for Route B. What summaries of these distributions would you use to select a route

   (a) to maximize the probability that you make it to work in less than 30 minutes?

   (b) to minimize your average commuting time?

4. For predictive distributions of survival time associated with two medical treatments, propose treatment selection criteria that are meaningful to you (or if you prefer, to society).

5. Here is an example in a vein similar to that of Example 1.2, and originally presented by Berger and Berry (1988). Consider a clinical trial established to study the effectiveness of Vitamin C in treating the common cold. After grouping subjects into pairs based on baseline variables such as gender, age, and health status, we randomly assign one member of each pair to receive Vitamin C, with the other receiving a placebo. We then count how many pairs had Vitamin C giving superior relief after 48 hours. We wish to test H_0 P(Vitamin C better) $= \frac{1}{2}$ versus H_a : P(Vitamin C better) $\neq \frac{1}{2}$.

   (a) Consider the expermental design wherein we sample $n = 17$ pairs, and observe $x = 13$ preferences for Vitamin C. What is the p-value for the above two-sided test?

   (b) Now consider a *two-stage* design, wherein we first sample $n_1 = 17$ pairs, and observe x_1 preferences for Vitamin C. In this design. if $x_1 \geq 13$ or $x_1 \leq 4$, we stop and reject H_0 Otherwise. we sample an *additional* $n_2 = 27$ pairs, and subsequently reject H_0 if $X_1 + X_2 \geq 29$ or $X_1 + X_2 \leq 15$. (This second stage rejection region was chosen since under H_0, $P(X_1 + X_2 \geq 29 \text{ or } X_1 + X_2 \leq 15) = P(X_1 \geq 13 \text{ or } X_1 \leq 4)$, the p-value in part (a) above.)

   If we once again observe $x_1 = 13$, what is the p-value under this new design? Is your answer consistent with that in part (a)?

   (c) What would be the impact on the p-value if we kept adding stages to the design, but kept observing $x_1 = 13$?

   (d) How would you analyze these data in the presence of a necessary but unforeseen change in the design - say, because the first 5 patients developed an allergic reaction to the treatment, and the trial was stopped by its clinicians.

   (e) What does all this suggest about the claim that p-values constitute "objective evidence" against H_0?

6. In the Normal/Normal example of Subsection 1.5.1, let $\sigma^2 = 2$, $\mu = 0$, and $\tau = 2$.

   (a) Suppose we observe $y = 4$. What are the mean and variance of the resulting posterior distribution? Sketch the prior, likelihood, and posterior on a single set of coordinate axes.

   (b) Repeat part (a) assuming $\tau^2 = 18$. Explain any resulting differences. Which of these two priors would likely have more appeal for a frequentist statistician?

7. In the basic diagnostic test setting, a disease is either present ($D = 1$) or absent ($D = 0$), and the test indicates either disease ($T = 1$) or no disease ($T = 0$). Represent $P(D = d|T = t)$ in terms of test sensitivity,

$P(T=1|D=1)$, specificity, $P(T=O|D=O)$, and disease prevalence, $P(D=1)$, and relate to Bayes' theorem (1.4).

8. In analyzing data from a Bin(n, θ) likelihood, the MLE is $\hat{\theta}_{MLE} = Y/n$, which has MSE = $E_{y|\theta}(\hat{\theta}_{MLE} - \theta)^2 = Var_{y|\theta}(\hat{\theta}_{MLE}) = \theta(1-\theta)/n$. Find the MSE of the estimator $\hat{\theta}_{Bayes} = (Y+1)/(n+2)$ and discuss in what contexts you would prefer it over $\hat{\theta}_{MLE}$. ($\hat{\theta}_{Bayes}$ is the estimator from equation (1.9) with $\mu = 1/2$ and $M = 2$.)

CHAPTER 2

The Bayes approach

2.1 Introduction

We begin by reviewing the fundamentals introduced in Chapter 1. In the Bayesian approach, in addition to specifying the model for the observed data $\mathbf{y} = y_1, \ldots, y_n$ given a vector of unknown parameters $\boldsymbol{\theta}$ usually in the form of a probability distribution $f(\mathbf{y}|\boldsymbol{\theta})$, we suppose that $\boldsymbol{\theta}$ is a random quantity as well, having a *prior* distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of *hyperparameters*. Inference concerning $\boldsymbol{\theta}$ is then based on its *posterior* distribution, given by

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\eta}) &= \frac{p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\eta})}{p(\mathbf{y}|\boldsymbol{\eta})} = \frac{p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\eta})}{\int p(\mathbf{y},\mathbf{u}|\boldsymbol{\eta})\,d\mathbf{u}} \\ &= \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta})\,d\mathbf{u}} . \end{aligned} \tag{2.1}$$

We refer to this formula as *Bayes' Theorem*. Notice the contribution of both the experimental data (in the form of the likelihood f) and prior opinion (in the form of the prior π) to the posterior in the last expression of equation (2.1). The result of the integral in the denominator is sometimes written as $m(\mathbf{y}|\boldsymbol{\eta})$, the *marginal* distribution of the data $\mathbf{y}$ given the value of the hyperparameter $\boldsymbol{\eta}$. If $\boldsymbol{\eta}$ is known, we will often suppress it in the notation, since there is no need to express conditioning on a constant. For example, the posterior distribution would be written simply as $p(\boldsymbol{\theta}|\mathbf{y})$.

Example 2.1 Consider the normal (Gaussian) likelihood, where $f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(y-\theta)^2}{2\sigma^2})$, $y \in \Re$, $\theta \in \Re$ and σ is a known positive constant. Henceforth we employ the shorthand notation $f(y|\theta) = N(y|\theta, \sigma^2)$ to denote a normal density with mean θ and variance σ^2. Suppose we take $\pi(\theta|\boldsymbol{\eta}) = N(\theta|\mu,\tau^2)$ where $\mu \in \Re$ and $\tau > 0$ are known hyperparameters, so that $\boldsymbol{\eta} = (\mu, \tau)$. By plugging these expressions into equation (2.1), it is fairly easy to show (Exercise 1 in Section 2.6) that the posterior distribution for θ is given by

$$p(\theta|y) = N\left(\theta \,\middle|\, \frac{\sigma^2\mu + \tau^2 y}{\sigma^2+\tau^2} \, , \, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) . \tag{2.2}$$

Writing $B = \frac{\sigma^2}{\sigma^2+\tau^2}$ we see that this posterior distribution has mean $B\mu + (1 - B)y$ and variance $B\tau^2 \equiv (1-B)\ \sigma^2$. Since $0 < B < 1$, the posterior mean is a weighted average of the prior mean and the observed data value, with weights that are inversely proportional to the corresponding variances. For this reason, B is sometimes called a *shrinkage factor,* since it gives the proportion of the distance that the posterior mean is "shrunk back" from the ordinary frequentist estimate y toward the prior mean μ. Note that when τ^2 is large relative to σ^2 (i.e., vague prior information) B is small and the posterior mean is close to the data value y. On the other hand, when τ^2 is small relative to σ^2 (i.e., a highly informative prior), B is large and the posterior mean is close to the prior mean μ.

Turning to the posterior variance, note that it is smaller than the variance of both the prior and the likelihood. In particular, it is easy to show (Exercise 1 in Section 2.6) that the precision in the posterior is the sum of the precisions in the likelihood and the prior, where *precision is* defined to be the reciprocal of the variance. Thus the posterior distribution offers a sensible compromise between our prior opinion and the observed data, and the combined strength of the two sources of information leads to increased precision in our understanding of θ ■

Given a sample of n independent observations we can obtain $f(\mathbf{y}|\boldsymbol{\theta})$ as $\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$ and proceed with equation (2.1). But evaluating this expression may be simpler if we can find a statistic $S(\mathbf{y})$ which is *sufficient* for $\boldsymbol{\theta}$ (that is, for which $f(\mathbf{y}|\boldsymbol{\theta} = h(\mathbf{y})g(S(\mathbf{y})|\boldsymbol{\theta}))$. To see this, note that for $S(\mathbf{y}) = s$,

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u})\,d\mathbf{u}} \\
&= \frac{h(\mathbf{y})g(S(\mathbf{y})|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int h(\mathbf{y})g(S(\mathbf{y})|\mathbf{u})\pi(\mathbf{u})\,d\mathbf{u}} \\
&= \frac{g(s|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(s)} \;=\; p(\boldsymbol{\theta}|s)\ ,
\end{aligned}
$$

provided $m(s) > 0$, since $h(\mathbf{y})$ cancels in the numerator and denominator of the middle expression. So if we can find a sufficient statistic, we may work with it instead of the entire dataset **y**, thus reducing the dimensionality of the problem.

Example 2.2 Consider again the normal/normal setting of Example 2.1, but where we now have an independent sample of size n from $f(\mathbf{y}|\ \theta)$. Since $S(\mathbf{y}) = \bar{y}$ is sufficient for θ we have that $p(\theta|\mathrm{v}) = \mathrm{v}(\theta|\bar{y})$. But. since we know that $f(\bar{y}|\theta) = N(\theta, \sigma^2/n)$, equation (2.2) implies that

$$
p(\theta|\bar{y}) \;=\; N\left(\theta \,\middle|\, \frac{(\sigma^2/n)\mu + \tau^2\bar{y}}{(\sigma^2/n) + \tau^2}\ ,\ \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2}\right)
$$

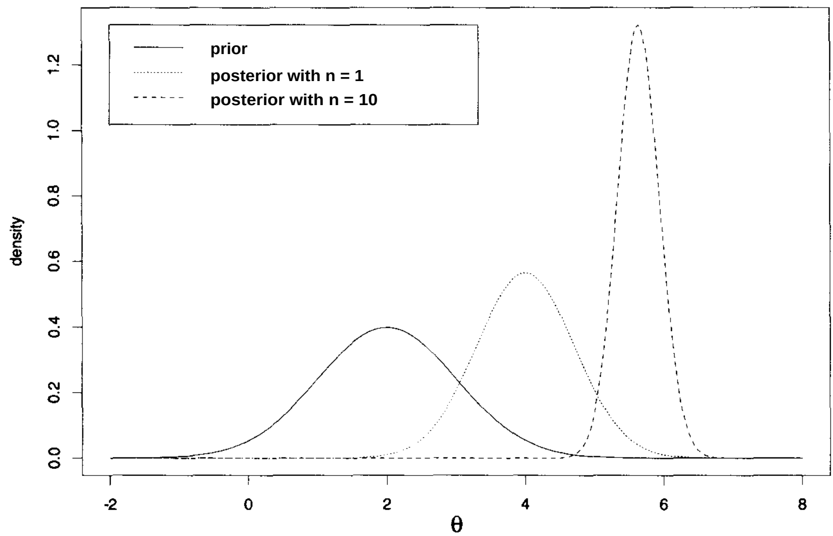


Figure 2.1 *Prior and posterior distributions based on samples of size 1 and 10, normal/normal model.*

$$= \quad N\left(\theta \,\middle|\, \frac{\sigma^2\mu + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right) .$$

As a concrete example, suppose that $\mu = 2$, $\tau = 1$, $\bar{y} = 6$, and $\sigma = 1$. Figure 2.1 plots the prior distribution, along with the posterior distributions arising from two different sample sizes. When n = 1, the prior and likelihood receive equal weight, and hence the posterior mean is $4 = \frac{2+6}{2}$. When $n = 10$, the data dominate the prior, resulting in a posterior mean much closer to $\bar{y}$. Notice that the posterior variance also shrinks as n gets larger; the posterior collapses to a point mass on $\bar{y}$ as n tends to infinity. ■

If we are unsure as to the proper value for $\boldsymbol{\eta}$, the proper Bayesian solution would be to quantify this uncertainty in a second-stage prior distribution (sometimes called a *hyperprior*). Denoting this distribution by $h(\boldsymbol{\eta})$, the desired posterior for $\boldsymbol{\theta}$ is now obtained by also marginalizing over $\boldsymbol{\eta}$,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int p(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\eta})\, d\boldsymbol{\eta}}{\int\int p(\mathbf{y},\mathbf{u},\boldsymbol{\eta})\, d\boldsymbol{\eta}\, d\mathbf{u}} \\ &= \frac{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})h(\boldsymbol{\eta})\, d\boldsymbol{\eta}}{\int\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta})h(\boldsymbol{\eta})\, d\boldsymbol{\eta}\, d\mathbf{u}} . \qquad (2.3) \end{aligned}$$

Alternatively, we might simply replace $\boldsymbol{\eta}$ by an estimate $\hat{\boldsymbol{\eta}}$ obtained as the value which maximizes the marginal distribution $m(\mathbf{y}|\boldsymbol{\eta})$ viewed as a function of $\boldsymbol{\eta}$. Inference is now based on the *estimated posterior* distribution $p(\theta|\mathbf{y}, \hat{\boldsymbol{\eta}})$, obtained by plugging $\hat{\boldsymbol{\eta}}$ into equation (2.1) above. Such an approach is often referred to as *empirical Bayes* analysis, since we are using the data to estimate the prior parameter $\boldsymbol{\eta}$. This is the subject of Chapter 3, where we also draw comparisons with the straight Bayesian approach using hyperpriors (c.f. Meeden, 1972; Deely and Lindley, 1981).

Of course, in principle, there is no reason why the hyperprior for $\boldsymbol{\eta}$ cannot itself depend on a collection of unknown parameters $\boldsymbol{\lambda}$, resulting in a generalization of (2.3) featuring a second-stage prior $h(\boldsymbol{\eta}|\boldsymbol{\lambda}$ and a third-stage prior $g(\boldsymbol{\lambda})$. This enterprise of specifying a model over several levels is called *hierarchical modeling,* with each new distribution forming a new level in the hierarchy. The proper number of levels varies with the problem, though typically our knowledge at levels above the second prior stage is sufficiently vague that additional levels are of little benefit. Moreover, since we are continually adding randomness as we move down the hierarchy, subtle changes to levels near the top are not likely to have much of an impact on the one at the bottom (the data level), which is typically the only one for which we actually have observations. In order to concentrate on more foundational issues, in this chapter we will typically limit ourselves to the simplest model having only two levels (likelihood and prior).

Continuing then with this two-level model (and again suppressing the dependence of the prior on the known value of $\boldsymbol{\eta}$), we observe that equation (2.1) may be expressed in the convenient shorthand

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ ,$$

or in words, "the posterior is proportional to the likelihood times the prior." Clearly the likelihood may be multiplied by any constant (or any function of $\mathbf{y}$ alone) without altering the posterior. Bayes Theorem may also be used sequentially: suppose we have two independently collected samples of data, $\mathbf{y}_1$ and $\mathbf{y}_2$. Then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2) &\propto f(\mathbf{y}_1, \mathbf{y}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= f_2(\mathbf{y}_2|\boldsymbol{\theta})f_1(\mathbf{y}_1|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto f_2(\mathbf{y}_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_1) \ . \end{aligned} \tag{2.4}$$

That is, we can obtain the posterior for the full dataset $\mathbf{y}_1, \mathbf{y}_2$ by first finding $p(\boldsymbol{\theta}|\mathbf{y}_1)$ and then treating it as the prior for the second portion of the data $\mathbf{y}_2$. This easy algorithm for updating the posterior is quite natural when the data arrive sequentially over time, as in a clinical trial or perhaps a business or economic setting.

Many authors (notably Geisser, 1993) have argued that concentrating on inference for the model parameters is misguided, since $\boldsymbol{\theta}$ is merely an unobservable, theoretical quantity. Switching to a different model for the data may result in an entirely different $\boldsymbol{\theta}$ vector. Moreover, even a perfect understanding of the model does not constitute a direct attack on the problem of predicting how the system under study will behave in the future - often the real goal of a statistical analysis. To this end, suppose that y_{n+1} is a future observation, independent of $\mathbf{y}$ given the underlying $\boldsymbol{\theta}$. Then the *predictive* distribution for y_{n+1} is given by

$$p(y_{n+1}|\mathbf{y}) = \int f(y_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}\ , \tag{2.5}$$

thanks to the conditional independence of y_{n+1} and y. This distribution summarizes the information concerning the likely value of a new observation, given the likelihood, the prior, and the data we have observed so far. (We remark that some authors refer to this distribution as the *posterior predictive,* and refer to the marginal distribution $m(y_{n+1}) = \int f(y_{n+1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ as the *prior predictive,* since the latter summarizes our information concerning y_{n+1} before having seen the data.)

The Bayesian decision-making paradigm improves on the traditional, frequentist approach to statistical analysis in its more philosophically sound foundation, its unified and streamlined approach to data analysis, and its ability to formally incorporate the prior opinion of one or more experimenters into the results via the prior distribution π. Practicing statisticians and biostatisticians, once reluctant to adopt the Bayesian approach due to general skepticism concerning its philosophy and a lack of necessary computational tools, are now turning to it with increasing regularity as traditional analytic approaches emerge as both theoretically and practically inadequate. For example, hierarchical Bayes methods form an ideal setting for combining information from several published studies of the same research area, an emerging scientific discipline commonly referred to as meta-analysis (DuMouchel and Harris, 1983; Cooper and Hedges, 1994). More generally, the hierarchical structure allows for honest assessment of heterogeneity both within and between groups such as laboratories or census areas. The review article by Breslow (1990) and the accompanying discussion provide an excellent summary of past and potential future application areas for Bayes and empirical Bayes methods in the public health and biomedical sciences.

For statistical models of even moderate complexity, the integrals appearing in (2.1) and (2.3) are not tractable in closed form, and thus must be evaluated numerically. Fortunately, this problem has been largely solved in recent years thanks to the development of methodology designed to take advantage of modern computing power. In particular, sampling-based methods for estimating hyperparameters and computing posterior distributions play a central role in modern Bayes and empirical Bayes analysis. Chapter 5 contains an extensive discussion of the various computational methods useful in Bayesian data analysis; for the time being, we assume that all relevant posterior distributions may be obtained, either analytically or numerically.

In the remainder of this chapter, we discuss the steps necessary to conduct a fully Bayesian data analysis, including prior specification, posterior and predictive inference, and diagnosis of departures from the assumed prior and likelihood form. Throughout our discussion, we illustrate with examples from the realms of biomedical science, public health, and environmental risk assessment as appropriate.

2.2 Prior distributions

Implementation of the Bayesian approach as indicated in the previous subsection depends on a willingness to assign probability distributions not only to data variables like **y**, but also to parameters like $\boldsymbol{\theta}$. Such a requirement may or may not be consistent with the usual long-run frequency notion of probability. For example, if

$$\theta = \text{true probability of success for a new surgical procedure,}$$

then it is possible (at least conceptually) to think of θ as the limiting value of the observed success rate as the procedure is independently repeated again and again. But if

$$\theta = \text{true proportion of U.S. men who are HIV-positive,}$$

the long-term frequency notion does not apply - it is not possible to even imagine "running the HIV epidemic over again" and reobserving θ. Moreover, the randomness in θ does not arise from any real-world mechanism; if an accurate census of all men and their HIV status were available, θ could be computed exactly. Rather, here θ is random only because it is unknown *to us,* though we may have some feelings about it (say, that $\theta = .05$ is more likely than $\theta = .50$). Bayesian analysis is predicated on such a belief in *subjective probability,* wherein we quantify whatever feelings (however vague) we may have about θ *before* we look at the data y in a distribution π. This distribution is then updated by the data via Bayes' Theorem (as in (2.1) or (2.3)) with the resulting posterior distribution reflecting a blend of the information in the data and the prior.

Historically, a major impediment to widespread use of the Bayesian paradigm has been that determination of the appropriate form of the prior π (and perhaps the hyperprior h) is often an arduous task. Typically, these distributions are specified based on information accumulated from past studies or from the opinions of subject-area experts. In order to streamline the elicitation process, as well as simplify the subsequent computational burden, experimenters often limit this choice somewhat by restricting π to some familiar distributional family. An even simpler alternative, available in some cases, is to endow the prior distribution with little informative content, so that the data from the current study will be the dominant force in determining the posterior distribution. We address each of these approaches in turn.

2.2.1 Elicited Priors

Suppose for the moment that $\boldsymbol{\theta}$ is univariate. Perhaps the simplest approach to specifying $\pi(\theta)$ is first to limit consideration to a manageable collection of θ values deemed "possible," and subsequently to assign probability masses to these values in such a way that their sum is 1, their relative contributions reflecting the experimenter's prior beliefs as closely as possible. If θ is discrete-valued, such an approach may be quite natural, though perhaps quite time-consuming. If θ is continuous, we must instead assign the masses to intervals on the real line, rather than to single points, resulting in a histogram prior for θ. Such a histogram (necessarily over a bounded region) may seem inappropriate, especially in concert with a continuous likelihood $f(\mathbf{y}|\theta)$ but may in fact be more appropriate if the integrals required to compute the posterior distribution must be done numerically, since computers are themselves discrete finite instruments. Moreover, a histogram prior may have as many bins as the patience of the elicitee and the accuracy of his prior opinion will allow. It is vitally important, however, that the range of the histogram be sufficiently wide, since, as can be seen from (2.1), the support of the posterior will necessarily be a subset of that of the prior. That is, the data cannot lend credence to intervals for θ deemed "impossible" in the prior.

Alternatively, we might simply assume that the prior for $\boldsymbol{\theta}$ belongs to a parametric distributional family $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$ choosing $\boldsymbol{\eta}$ so that the result matches the elicitee's true prior beliefs as nearly as possible. For example, if $\boldsymbol{\eta}$ were two-dimensional, then specification of two moments (say, the mean and the variance) or two quantiles (say, the 50th and 95th) would be sufficient to determine its exact value. This approach limits the effort required of the elicitee, and also overcomes the finite support problem inherent in the histogram approach. It may also lead to simplifications in the posterior computation, as we shall see in Subsection 2.2.2.

A limitation of this approach is of course that it may not be possible for the elicitee to "shoehorn" his or her prior beliefs into any of the standard parametric forms. In addition, two distributions which look virtually identical may in fact have quite different properties. For example, Berger (1985, p. 79) points out that the Cauchy(0,1) and Normal(0, 2.19) distributions have identical 25th, 50th, and 75th percentiles (-1, 0, and l, respectively) and density functions that appear very similar when plotted, yet may lead to quite different posterior distributions.

An opportunity to experiment with these two types of prior elicitation in a simple setting is afforded below in Exercise 4. We now illustrate multivariate elicitation in a challenging, real-life setting.

Example 2.3 In attempting to model the probability p_{ijkl} of an incorrect response by person *j* in cue group *i* on a working memory exam question having *k* "simple conditions" and "query complexity" level *l*, Carlin, Kass, Lerch, and Huguenard (1992) consider the model

$$logit(p_{ijkl}) = \begin{cases} \theta_j^{(i)} + \gamma k, & l = 0 \\ \theta_j^{(i)} + \gamma k + \alpha, & l = 1 \\ \theta_j^{(i)} + \gamma k + \alpha + \beta, & l = 2 \end{cases} .$$

In this model, $\theta_j^{(i)}$ denotes the effect for subject *ij*, γ denotes the effect due to the number of simple conditions in the question, α measures the marginal increase in the logit error rate in moving from complexity level 0 to 1, and β does the same in moving from complexity level 1 to 2. Prior information concerning the parameters was based primarily on rough impressions, precluding precise prior elicitation. As such, $\pi(\{\theta_j^{(i)}\}, \gamma, \alpha, \beta)$ was defined as a product of independent, normally distributed components, namely,

$$\prod_{i=1}^{I} \prod_{j=1}^{J} N(\theta_j^{(i)} | \mu_\theta, \sigma_\theta^2) \times N(\gamma | \mu_\gamma, \sigma_\gamma^2) \times N_2(\alpha, \beta | \mu_\alpha, \mu_\beta, \sigma_\alpha^2, \sigma_\beta^2, \rho) \qquad (2.6)$$

where N represents the univariate normal distribution and N_2 represents the bivariate normal distribution. Hyperparameter mean parameters were chosen by eliciting "most likely" error rates, and subsequently converting to the logit scale. Next, variances were determined by considering the effect on the error rate scale of a postulated standard deviation (and resulting 95% confidence set) on the logit scale. Note that the symmetry imposed by the normal priors on the logit scale does not carry over to the error rate (probability) scale. While neither normality nor prior independence was deemed totally realistic, prior (2.6) was able to provide a suitable starting place, subsequently updated by the information in the experimental data. ■

Even when scientifically relevant prior information is available, elicitation of the precise forms for the prior distribution from the experimenters can be a long and tedious process. Some progress has been made in this area for normal linear models, most notably by Kadane et al. (1980). In the main, however, prior elicitation issues tend to be application- and elicitee-specific, meaning that general purpose algorithms are typically unavailable. Wolfson (1995) and Chaloner (1996) provide overviews of the various philosophies of elicitation (based on the ways that humans think about and update probabilistic statements), as well as reviews of the methods proposed for various models and distributional settings.

The difficulty of prior elicitation has been ameliorated somewhat through the addition of interactive computing, especially dynamic graphics, and object-oriented computer languages such as S (Becker, Chambers, and Wilks, 1988) or XLISP-STAT (Tierney, 1990).

Example 2.4 In the arena of monitoring clinical trials, Chaloner et al. (1993) show how to combine histogram elicitation, matching a functional form, and interactive graphical methods. Following the advice of Kadane et al. (1980), these authors elicit a prior not on θ (in this case, an unobservable proportional hazards regression parameter), but on corresponding observable quantities familiar to their medically-oriented elicitees, namely the proportion of individuals failing within two years in a population of control patients, p_0 and the corresponding two-year failure proportion in a population of treated patients, p_1. Writing the survivor function at time t for the controls as $S(t)$, under the proportional hazards model we then have that $p_0 = 1 - S(2)$ and $p_1 = 1 - S(2)^{\exp(\theta)}$ Hence the equation

$$\log[-\log(1-p_1)] = \theta + \log[-\log(1-p_0)] \tag{2.7}$$

gives the relationship between p_0, p_1 and θ.

Since the effect of a treatment is typically thought of by clinicians as being relative to the baseline (control) failure probability p_0 they first elicit a best guess for this rate, $\hat{p}_0$. Conditional on this modal value, they then elicit an entire distribution for p_1, beginning with initial guesses for the upper and lower quartiles of p_1's distribution. These values determine initial guesses for the parameters μ and σ of a smooth density function that corresponds on the θ scale to an extreme value distribution. This smooth functional form is then displayed on the computer screen and the elicitee is allowed to experiment with new values for μ and σ in order to obtain an even better fit to his or her true prior beliefs. Finally, fine tuning of the density is allowed via mouse input directly onto the screen. The density is restandardized after each such change, with updated quantiles computed for the elicitee's approval. At the conclusion of this process, the final prior distribution is discretized onto a suitably fine grid of points $(p_{11}, \ldots, p_{1K})$, and finally converted to a histogram-type prior on θ via equation (2.7), for use in computing its posterior distribution. ■

2.2.2 Conjugate priors

In choosing a prior belonging to a specific distributional family $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$ some choices may be more computationally convenient than others. In particular, it may be possible to select a member of that family which is *conjugate* to the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$, that is, one that leads to a posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ belonging to the same distributional family as the prior. Morris (1983b) showed that exponential families, from which we typically draw our likelihood functions, do in fact have conjugate priors, so that this approach will often be available in practice. The computational advantages are best illustrated through an example.

Example 2.5 Suppose that X is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest adopting a Poisson likelihood,

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},\ x \in \{0,1,2,\ldots\},\ \theta > 0.$$

To effect a Bayesian analysis we require a prior distribution for θ having support on the positive real line. A reasonably flexible choice is provided by the gamma distribution,

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha},\ \theta > 0, \alpha > 0,\ \beta > 0,$$

or $\theta \sim G(\alpha, \beta)$ in distributional shorthand. Note that we have suppressed π's dependence on $\boldsymbol{\eta} = (\alpha, \beta)$ since we assume it to be known. The gamma distribution has mean $\alpha\beta$ variance $\alpha\beta^2$, and can have a shape that is either one-tailed ($\alpha \leq 1$) or two-tailed ($\alpha > 1$). (The scale of the distribution is affected by β but its shape is not.) Using Bayes' Theorem (2.1) to obtain the posterior density, we have

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)\pi(\theta) \\ &\propto \left(e^{-\theta}\theta^x\right)\left(\theta^{\alpha-1}e^{-\theta/\beta}\right) \\ &= \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)}\ . \end{aligned} \tag{2.8}$$

Notice that since our intended result is a normalized function of θ we are able to drop any multiplicative functions that do not depend on θ. (For example, in the first line we have dropped the marginal distribution $m(x)$ in the denominator, since it is free of θ.) But now looking at (2.8), we see that it is proportional to a gamma distribution with parameters $\alpha' = x + \alpha$ and $\beta' = (1 + 1/\beta)^{-1}$. Since this is the only function proportional to (2.8) that integrates to 1 and density functions uniquely determine distributions, we know that the posterior distribution for θ is indeed $G(\alpha', \beta')$, and that the gamma is the conjugate family for the Poisson likelihood. ■

In this example, we might have guessed the identity of the conjugate prior by looking at the Poisson likelihood as a function of θ instead of x as we typically do. Such a viewpoint reveals the $\theta^a e^{-b\theta}$ structure that is also the basis of the gamma density. This technique is useful for determining conjugate priors in many exponential families, as we shall see as we continue. Since conjugate priors can permit posterior distributions to emerge without numerical integration, we shall use them in many of our subsequent examples, in order to concentrate on foundational concerns and defer messy computational issues until Chapter 5.

For multiparameter models, independent conjugate priors may often be specified for each parameter, leading to corresponding conjugate forms for each *conditional* posterior distribution.

Example 2.6 Consider once again the normal likelihood of Example 2.1, where now both θ and σ^2 are assumed unknown, so that $\boldsymbol{\theta} = (\theta, \sigma^2)$. Looking at f as a function of θ alone, we see an expression proportional to $\exp(-(\theta - a)^2/b)$, so that the conjugate prior is again given by the normal distribution. Thus we assume that $\pi_1(\theta) = N(\theta|\mu, \tau^2)$ as before. Next, looking at f as a function of σ, we see a form proportional to $(\sigma^2)^{-a} e^{-b/\sigma^2}$ which is vaguely reminiscent of the gamma distribution. In fact, this is actually the reciprocal of a gamma: if $X \sim G(\alpha, \beta)$, then $Y = 1/X \sim$ Inverse Gamma$(\alpha, \beta) \equiv IG(\alpha, \beta)$, with density function (see Appendix Section A.2)

$$g(y|\alpha, \beta) = \frac{e^{-1/\beta y}}{\Gamma(\alpha)\beta^\alpha y^{\alpha+1}}, \;\; y > 0, \; \alpha > 0, \; \beta > 0.$$

Thus we assume that $\pi_2(\sigma^2) = IG(\sigma^2|a, b)$, where a and b are the shape and scale parameters, respectively. Finally, we assume that θ and σ^2 are *a priori* independent, so that $\pi(\boldsymbol{\theta}) = \pi_1(\theta)\pi_2(\sigma^2)$. With this componentwise conjugate prior specification, the two conditional posterior distributions emerge as expression (2.2) and

$$p(\sigma^2|x, \theta) = IG\left(\sigma^2 \;\Big|\; \frac{1}{2} + a\,, \; \left[\frac{1}{2}(x - \theta)^2 + \frac{1}{b}\right]^{-1}\right). \tag{2.9}$$

■

We remark that in the normal setting, closed forms may also emerge for the *marginal* posterior densities $p(\theta|x)$ and $p(\sigma^2|x)$; we leave the details as an exercise. In any case, the ability of conjugate priors to produce at least unidimensional conditional posteriors in closed form enables them to retain their importance even in very high dimensional settings. This occurs through the use of Markov chain Monte Carlo integration techniques, which construct a sample from the joint posterior by successively sampling from the individual conditional posterior distributions.

Finally, while a single conjugate prior may be inadequate to accurately reflect available prior knowledge, a finite mixture of conjugate priors may be sufficiently flexible (allowing multimodality, heavier tails, etc.) while still enabling simplified posterior calculations. For example, suppose we are given the likelihood $f(x|\theta)$ for the univariate parameter θ, and adopt the two-component mixture prior $\pi(\theta) = \alpha\pi_1(\theta) + (1 - \alpha)\pi_2(\theta)$, $0 \le \alpha \le 1$, where both π_1 and π_2 belong to the family of priors conjugate for f. Then

$$p(\theta|x) \;\; = \;\; \frac{f(x|\theta)\pi_1(\theta)\alpha + f(x|\theta)\pi_2(\theta)(1 - \alpha)}{\int [f(x|\theta)\pi_1(\theta)\alpha + f(x|\theta)\pi_2(\theta)(1 - \alpha)] d\theta}$$

$$
\begin{aligned}
&= \frac{\frac{f(x|\theta)\pi_1(\theta)}{m_1(x)} m_1(x)\alpha + \frac{f(x|\theta)\pi_2(\theta)}{m_2(x)} m_2(x)(1-\alpha)}{m_1(x)\alpha + m_2(x)(1-\alpha)} \\
&= \frac{f(x|\theta)\pi_1(\theta)}{m_1(x)} w_1 + \frac{f(x|\theta)\pi_2(\theta)}{m_2(x)} w_2 \\
&= p_1(\theta|x)w_1 + p_2(\theta|x)w_2 \ ,
\end{aligned}
$$

where $w_1 = m_1(x)\alpha/[m_1(x)\alpha + m_2(x)(1-\alpha)]$ and $w_2 = 1 - w_1$. Hence a mixture of conjugate priors leads to a mixture of conjugate posteriors.

2.2.3 Noninformative Priors

As alluded to earlier, often no reliable prior information concerning $\boldsymbol{\theta}$ exists, or an inference based solely on the data is desired. At first, it might appear that Bayesian inference would be inappropriate in such settings, but this conclusion is a bit hasty. Suppose we could find a distribution $\pi(\boldsymbol{\theta})$ that contained "no information" about $\boldsymbol{\theta}$ in the sense that it did not favor one $\boldsymbol{\theta}$ value over another (provided both values were logically possible). We might refer to such a distribution as a *noninformative prior* for $\boldsymbol{\theta}$, and argue that all of the information resulting in the posterior $p(\boldsymbol{\theta}|\mathbf{x})$ arose from the data, and hence all resulting inferences were completely *objective,* rather than subjective. Such an approach is likely to be important if Bayesian methods are to compete successfully in practice with their popular likelihood-based counterparts (e.g., maximum likelihood estimation). But is such a "noninformative approach" possible?

In some cases, the answer is an unambiguous "yes." For example, suppose the parameter space is discrete and finite, i.e., $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\}$. Then clearly the distribution

$$p(\theta_i) = 1/n, \ i = 1, \ldots, n,$$

does not favor any one of the candidate θ values over any other, and, as such, is noninformative for θ. If instead we have a bounded continuous parameter space, say $\boldsymbol{\Theta} = [a, b]$, $-\infty < a < b < \infty$, then the uniform distribution

$$p(\theta) = 1/(b-a), \ a < \theta < b,$$

is arguably noninformative for θ (though this conclusion can be questioned, as we shall see later in this subsection and in the example of Subsection 2.3.4).

When we move to unbounded parameter spaces, the situation is even less clear. Suppose that $\boldsymbol{\Theta} = (-\infty, \infty)$. Then the appropriate uniform prior would appear to be

$$p(\theta) = c, \ \text{any } c > 0.$$

But this distribution is *improper,* in that $\int p(\theta)d\theta = \infty$, and hence does not appear appropriate for use as a prior. But even here, Bayesian inference is still possible if the integral with respect to θ of the likelihood $f(\mathbf{x}|\theta)$ equals some finite value K. (Note that this will not necessarily be the case: the definition of f as a joint probability density requires only the finiteness of its integral with respect to $\mathbf{x}$.) Then

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta) \cdot c}{\int f(\mathbf{x}|\theta) \cdot c \, d\theta} = \frac{f(\mathbf{x}|\theta)}{K} \, .$$

Since the integral of this function with respect to θ is 1, it is indeed a proper posterior density, and hence Bayesian inference may proceed as usual. Still, it is worth reemphasizing that care must be taken when using improper priors, since proper posteriors will not always result. A good example (to which we return in Chapter 5 in the context of computation for high-dimensional models) is given by random effects models for longitudinal data, where the dimension of the parameter space increases with the sample size. In such models, the information in the data may be insufficient to identify all the parameters and, consequently, at least some of the prior distributions on the individual parameters must be informative.

Noninformative priors are closely related to the notion of a *reference prior*, though precisely what is meant by this term depends on the author. Kass and Wasserman (1996) define such a prior as a conventional or "default" choice - not necessarily noninformative, but a convenient place to begin the analysis. The definition of Bernardo (1979) stays closer to non-informativity, using an expected utility approach to obtain a prior which represents ignorance about θ in some formal sense. Unfortunately, using this approach (and others that are similar), a unique representation for the reference prior is not always possible. Box and Tiao (1973, p. 23) suggest that all that is important is that the data dominate whatever information is contained in the prior, since as long as this happens, the precise form of the prior is unimportant. They use the analogy of a properly run jury trial, wherein the weight of evidence presented in court dominates whatever prior ideas or biases may be held by the members of the jury.

Putting aside these subtle differences in definition for a moment, we ask the question of whether a uniform prior is always a good "reference." Sadly, the answer is no, since it is not invariant under reparametrization. As a simple example, suppose we claim ignorance concerning a univariate parameter θ defined on the positive real line, and adopt the prior p(θ 1, $\theta > 0$. A sensible reparametrization is given by $\gamma = g(\theta) = \log(\theta)$, since this converts the support of the parameter to the whole real line. Then the prior on γ is given by $p'(\gamma) = |J| \, p(e^{\gamma})$, where $J = d\theta/d\gamma$, the Jacobian of the inverse transformation. Here we have $\theta = g^{-1}(\gamma) = e^{\gamma}$, so that $J = e^{\gamma}$ and hence

$$p'(\gamma) = e^{\gamma}, \; -\infty < \gamma < \infty \, ,$$

a prior that is clearly *not* uniform. Thus, using uniformity as a universal definition of prior ignorance, it is possible that "ignorance about θ" does not imply "ignorance about γ!" A possible remedy to this problem is to rely on the particular modeling context to provide the most reasonable parametrization and, subsequently, apply the uniform prior on this scale.

The *Jeffreys* prior (Jeffreys, 1961, p.181) offers a fairly easy-to-compute alternative that is invariant under transformation. In the univariate case, this prior is given by

$$p(\theta) \propto [I(\theta)]^{1/2} , \tag{2.10}$$

where $I(\theta)$ is the expected Fisher information in the model, namely,

$$I(\theta) = -E_{\mathbf{x}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)\right] . \tag{2.11}$$

Notice that the *form* of the likelihood helps to determine the prior (2.10), but not the actual observed *data* since we are averaging over $\mathbf{x}$ in equation (2.11).

It is left as an exercise to show that the Jeffreys prior is invariant to 1-1 transformation. That is, in the notation of the previous paragraph, it is true that

$$[I(\gamma)]^{1/2} = [I(\theta)]^{1/2}\left|\frac{d\theta}{d\gamma}\right| , \tag{2.12}$$

so that computing the Jeffreys prior for γ directly produces the same answer as computing the Jeffreys prior for θ and subsequently performing the usual Jacobian transformation to the γ scale.

In the multiparameter case, the Jeffreys prior is given by

$$p(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2} , \tag{2.13}$$

where $|\cdot|$ now denotes the determinant, and $I(\boldsymbol{\theta})$ is the expected Fisher information *matrix*, having *ij*-element

$$I_{ij}(\boldsymbol{\theta}) = -E_{\mathbf{x}|\theta}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{x}|\boldsymbol{\theta})\right] .$$

While (2.13) provides a general recipe for obtaining noninformative priors, it can be cumbersome to use in high dimensions. A more common approach is to obtain a noninformative prior for each parameter individually, and then form the joint prior simply as the product of these individual priors. This action is often justified on the grounds that "ignorance" is consistent with "independence," although since noninformative priors are often improper, the formal notion of independence does not really apply.

We close this subsection with two important special cases often encountered in Bayesian modeling. Suppose the density of X is such that $f(x|\theta) = f(x-\theta)$. That is, the density involves θ only through the term $(x-\theta)$. Then θ is called a *location parameter*, and f is referred to as a *location parameter family*. Examples of location parameters include the normal mean and Cauchy median parameters. Berger (1985, pp.83–84) shows that, for a prior on θ to be invariant under location transformations (i.e., transformations of the form $Y = X + c$), it must be uniform over the range of θ. Hence the noninformative prior for a location parameter is given by

$$p(\theta) = 1, \ \theta \in \Re \ .$$

Next, assume instead that the density of X is of the form $f(x|\sigma) = \frac{1}{\sigma} f(\frac{x}{\sigma})$, where $\sigma > 0$. Then σ is referred to as a scale parameter, and f is a scale parameter family. Examples of scale parameters include the normal standard deviation (as our notation suggests), as well as the σ parameters in the Cauchy and t distributions and the β parameter in the gamma distribution. If we wish to specify a prior for σ that is invariant under scale transformations (i.e., transformations of the form $Y = cX$, for $c > 0$), then Berger shows we need only consider priors of the form k/σ, for $k > 0$. Hence the noninformative prior for a scale parameter is given by

$$p(\sigma) = \frac{1}{\sigma}, \ \sigma > 0 \ .$$

If we prefer to work on the variance (σ^2) scale, we of course obtain $p(\sigma^2) = 1/\sigma^2$. Note that, like the uniform prior for the location parameter, both of these priors are also improper since $\int_0^\infty d\sigma/\sigma = \int_0^\infty d\sigma^2/\sigma^2 = \infty$.

Finally, many of the most popular distributional forms (including the normal, t, and Cauchy) combine both location and scale features into the form $f(x|\theta, \sigma) = \frac{1}{\sigma} f(\frac{x-\theta}{\sigma})$, the so-called location-scale family. The most common approach in this case is to use the two noninformative priors given above in concert with the notion of prior "independence," obtaining

$$p(\theta, \sigma) = \frac{1}{\sigma}, \ \theta \in \Re, \ \sigma > 0 \ .$$

As discussed in the exercises, the Jeffreys prior for this situation takes a slightly different form.

2.2.4 *Other prior construction methods*

There are a multitude of other methods for constructing priors, both informative and noninformative; see Berger (1985, Chapter 3) for an overview. The only one we shall mention is that of using the marginal distribution $m(\mathbf{x}) = \int f(\mathbf{x}|\theta)p(\theta)d\theta$. Here, we choose the prior $p(\theta)$ based on its ability to preserve consistency with the marginal distribution of the observed data. Unlike the previous approaches in this section, this approach uses not only the form of the likelihood, but also the actual observed data values to help determine the prior. We might refer to this method as empirical estimation of the prior, and this is, in fact, the method that is used in empirical Bayes (EB) analysis.

Strictly speaking, empirical estimation of the prior is a violation of

Bayesian philosophy: the subsequent prior-to-posterior updating in equation (2.1) would "use the data twice" (first in the prior, and again in the likelihood). The resulting inferences from this posterior would thus be "overconfident." Indeed, EB methods that ignore this fact are often referred to as *naive* EB methods. However, much recent work focuses on ways to correct these methods for the effect of their estimation of the prior - many with a good deal of success. This has enabled EB methods to retain their popularity among practitioners despite the onslaught of advanced computational methods designed to tackle the complex integration present in (2.3). These EB methods are the subject of Chapter 3.

2.3 Bayesian inference

Having specified the prior distribution using one of the techniques in the previous section, we may now use Bayes' Theorem (2.1) to obtain the posterior distribution of the model parameters, or perhaps equation (2.5) to obtain the predictive density of a future observation. Since these distributions summarize our current state of knowledge (arising from both the observed data and our prior opinion, if any), we might simply graph the corresponding density (or cumulative distribution) functions and report them as the basis for all posterior inference. However, these functions can be difficult to interpret and, in many cases, may tell us more than we want to know. Hence in this section, we discuss common approaches for summarizing such distributions. In particular, we develop Bayesian analogues for the common frequentist techniques of point estimation, interval estimation, and hypothesis testing. Throughout, we work with the posterior distribution of the model parameters, though most of our methods apply equally well to the predictive distribution - after all, future data values and parameters are merely different types of unknown quantities. We illustrate how Bayesian data analysis is typically practiced, and, in the cases of interval estimation and hypothesis testing, the philosophical advantages inherent in the Bayesian approach.

2.3.1 *Point estimation*

For ease of notation, consider first the univariate case. To obtain a point estimate $\hat{\theta}(y)$ of θ, we simply need to select a summary feature of $p(\theta|\mathbf{y})$, such as its mean, median, or mode. The mode will typically be the easiest to compute, since no standardization of the posterior is then required: we may work directly with the numerator of (2.1). Also, note that when the prior $\pi(\theta)$ is flat, the posterior mode will be equal to the maximum likelihood estimate of θ. For this reason, the posterior mode is sometimes referred to as the *generalized maximum likelihood estimate* of θ.

For symmetric posterior densities, the mean and the median will of course

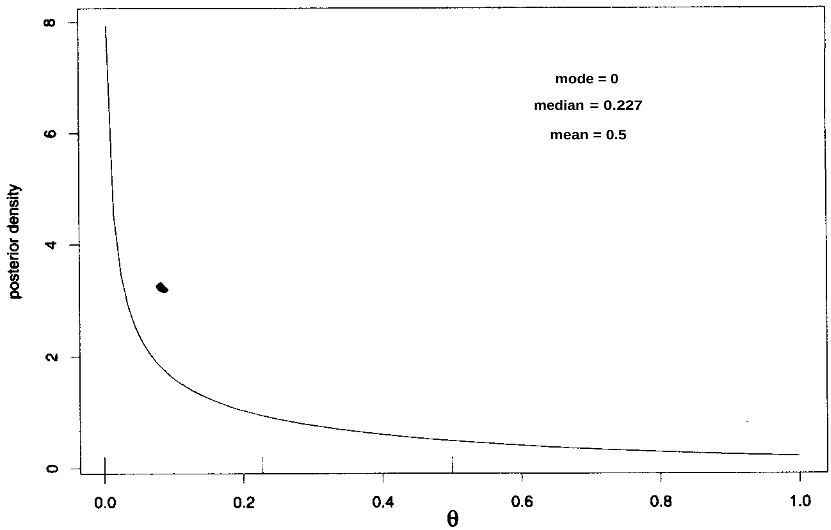


Figure 2.2 *Three point estimates arising from a Gamma(. 5, 1) posterior.*

be identical; for symmetric unimodal posteriors, all three measures will coincide. For asymmetric posteriors, the choice is less clear, though the median is often preferred since it is intermediate to the mode (which considers only the value corresponding to the maximum value of the density) and the mean (which often gives too much weight to extreme outliers). These differences are especially acute for one-tailed densities. Figure 2.2 provides an illustration using a $G(.5, 1)$ distribution. Here, the modal value (0) seems a particularly poor choice of summary statistic, since it is far from the "middle" of the density; in fact, the density is not even finite at this value.

In order to obtain a measure of the accuracy of a point estimate $\hat{\theta}(\mathbf{y})$ of θ, we might use the *posterior variance with respect to* $\hat{\theta}(\mathbf{y})$,

$$E_{\theta|\mathbf{y}}(\theta - \hat{\theta}(\mathbf{y}))^2 .$$

Writing the posterior mean $E_{\theta|\mathbf{y}}(\theta) \equiv \mu(\mathbf{y})$ simply as μ for the moment, we have

$$\begin{aligned} E_{\theta|\mathbf{y}}(\theta - \hat{\theta}(\mathbf{y}))^2 &= E_{\theta|\mathbf{y}}(\theta - \mu + \mu - \hat{\theta}(\mathbf{y}))^2 \\ &= E_{\theta|\mathbf{y}}[(\theta - \mu)^2 + 2(\theta - \mu)(\mu - \hat{\theta}(\mathbf{y})) + (\mu - \hat{\theta}(\mathbf{y}))^2] \\ &= Var_{\theta|\mathbf{y}}(\theta) + 2[\mu - \hat{\theta}(\mathbf{y})][E_{\theta|\mathbf{y}}(\theta) - \mu] + (\mu - \hat{\theta}(\mathbf{y}))^2 \\ &= Var_{\theta|\mathbf{y}}(\theta) + (\mu - \hat{\theta}(\mathbf{y}))^2 , \end{aligned}$$

since the middle term on the third line is identically zero. Hence we have shown that the posterior mean μ minimizes the posterior variance with respect to $\hat{\theta}(\mathbf{y})$ over all point estimators $\hat{\theta}(\mathbf{y})$. Furthermore, this minimum value is $Var_{\theta|\mathbf{y}}(\theta) = E_{\theta|\mathbf{y}}[\theta - E_{\theta|\mathbf{y}}(\theta)]^2$ the posterior variance of θ (usually referred to simply as the posterior variance). This is a possible argument for preferring the posterior mean as a point estimate, and also partially explains why historically only the posterior mean and variance were reported as the results of a Bayesian analysis.

Turning briefly to the multivariate case, we might again take the posterior mode as our point estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ though it will now be numerically harder to find. When the mode exists, a traditional maximization method such as a grid search, golden section search, or Newton-type method will often succeed in locating it (see e.g., Thisted, 1988, Chapter 4 for a description of these and other maximization methods). The posterior mean $\boldsymbol{\mu} \equiv E_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta})$ is also a possibility, since it is still well-defined and its accuracy is captured nicely by the posterior covariance matrix,

$$V = E_{\boldsymbol{\theta}|\mathbf{y}}\left[(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})'\right] .$$

Similar to the univariate case, one can show that

$$E_{\boldsymbol{\theta}|\mathbf{y}}\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{y}))(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{y}))'\right] = V + (\boldsymbol{\mu} - \hat{\boldsymbol{\theta}}(\mathbf{y}))(\boldsymbol{\mu} - \hat{\boldsymbol{\theta}}(\mathbf{y}))' ,$$

so that the posterior mean "minimizes" the posterior covariance matrix with respect to $\hat{\boldsymbol{\theta}}(\mathbf{y})$. The posterior median is more difficult to define in multiple dimensions, though it is still often used in one-dimensional subspaces of such problems, i.e., as a summary statistic for $p(\theta_1|\mathbf{y}) = \int \cdots \int p(\boldsymbol{\theta}|\mathbf{y})\, d\theta_2, \ldots, d\theta_k$.

Example 2.7 Perhaps the single most important contribution of statistics to the field of scientific inquiry is the general linear model, of which regression and analysis of variance models are extremely widely-used special cases. A Bayesian analysis of this model was first presented in the landmark paper by Lindley and Smith (1972), which we summarize here. Suppose that $\mathbf{Y}|\boldsymbol{\theta}_1 \sim N(A_1\boldsymbol{\theta}_1, C_1)$ where $\mathbf{Y}$ is an $n \times 1$ data vector, $\boldsymbol{\theta}_1$ is a $p_1 \times 1$ parameter vector, A_1 is an $n \times p_1$ known design matrix, and C_1 is an $n \times n$ known covariance matrix. Suppose further that we adopt the prior distribution $\boldsymbol{\theta}_1 \sim N(A_2\boldsymbol{\theta}_2, C_2)$, where $\boldsymbol{\theta}_2$ is a $p_2 \times 1$ parameter vector, A_2 is a $p_1 \times p_2$ design matrix, C_2 is a $p_1 \times p_1$ covariance matrix, and $\boldsymbol{\theta}_2, A_2$, and C_2 are all known. Then the marginal distribution of $\mathbf{Y}$ is

$$\mathbf{Y} \sim N(A_1A_2\boldsymbol{\theta}_2 \,,\, C_1 + A_1C_2A_1') ,$$

and the posterior distribution of $\boldsymbol{\theta}_1$ is

$$\boldsymbol{\theta}_1|\mathbf{y} \sim N(D\mathbf{d} \,,\, D) , \tag{2.14}$$

where

$$D^{-1} = A_1' C_1^{-1} A_1 + C_2^{-1} , \qquad (2.15)$$

and

$$\mathbf{d} = A_1' C_1^{-1} \mathbf{y} + C_2^{-1} A_2 \boldsymbol{\theta}_2 . \qquad (2.16)$$

Thus $E(\boldsymbol{\theta}_1|\mathbf{y}) = D\mathbf{d}$ provides a point estimate for $\boldsymbol{\theta}_1$ with associated variability captured by the posterior covariance matrix Var$(\boldsymbol{\theta}_1|\mathbf{y}) = D$.

As a more concrete illustration, we consider the case of linear regression. Using the usual notation, we would set $A_1 = X$, $\theta_1 = \boldsymbol{\beta}$, and $C_1 = \sigma^2 I_n$. A noninformative prior is provided by taking $C_2^{-1} = \mathbf{0}$, i.e., setting the prior precision matrix equal to a $p_1 \times p_1$ matrix of zeroes. Then from equations (2.15) and (2.16), we have

$$D^{-1} = X'(\sigma^2 I_n)^{-1} X + \mathbf{0} = \frac{1}{\sigma^2}(X'X) ,$$

and

$$\mathbf{d} = X'(\sigma^2 I_n)^{-1} \mathbf{y} + \mathbf{0} = \frac{1}{\sigma^2}(X'\mathbf{y}) ,$$

so that the posterior mean is given by

$$D\mathbf{d} = \left[\frac{1}{\sigma^2}(X'X)\right]^{-1} \frac{1}{\sigma^2}(X'\mathbf{y}) = (X'X)^{-1}X'\mathbf{y} = \hat{\boldsymbol{\beta}}_{LS} ,$$

the usual least squares estimate of $\boldsymbol{\beta}$. From (2.14), the posterior distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}_{LS}, \sigma^2(X'X)^{-1}) .$$

Recall that the sampling distribution of the least squares estimate is given by $\hat{\boldsymbol{\beta}}_{LS}|\boldsymbol{\beta} \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$ so that classical and noninformative Bayesian inferences regarding $\boldsymbol{\beta}$ will be formally identical in this example.

If $C_1 = Var(\mathbf{Y}|\boldsymbol{\theta}_1)$ is unknown, a closed form analytic solution for the posterior mean of $\boldsymbol{\beta}$ will typically be unavailable. Fortunately, sampling-based computational methods are available to escape this unpleasant situation, as we shall illustrate later in Example 5.6. ■

2.3.2 Interval estimation

The Bayesian analogue of a frequentist confidence interval (CI) is usually referred to as a *credible set,* though we will often use the term "Bayesian confidence interval" or simply "confidence interval" in univariate cases for consistency and clarity. More formally,

Definition 2.1 A 100 x (1 - α)% credible set for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ such that

$$1 - \alpha \le P(C|\mathbf{y}) = \int_C p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} ,$$

where integration is replaced by summation for discrete components of $\boldsymbol{\theta}$.

This definition enables direct probability statements about the likelihood of $\boldsymbol{\theta}$ falling in C, i.e.,

"The probability that θ lies in C given the observed data $\mathbf{y}$ is at least $(1-\alpha)$."

This is in stark contrast to the usual frequentist CI, for which the corresponding statement would be something like,

"If we could recompute C for a large number of datasets collected in the same way as ours, about $(1 - \alpha) \times 100\%$ of them would contain the true value of θ."

This is not a very comforting statement, since we may not be able to even imagine repeating our experiment a large number of times (e.g., consider an interval estimate for the 1993 U.S. unemployment rate). In any event, we are in physical possession of only one dataset; our computed C will either contain $\boldsymbol{\theta}$ or it won't, so the actual coverage probability will be either 1 or 0. Thus for the frequentist, the confidence level $(1- \alpha$ is only a "tag" that indicates the quality of the procedure (i.e., a narrow 95% CI should make us feel better than an equally narrow 90% one). But for the Bayesian, the credible set provides an actual probability statement, based only on the observed data and whatever prior opinion we have added.

We used "$\leq$" instead of "=" in Definition 2.1 in order to accommodate discrete settings, where obtaining an interval with coverage probability *exactly* $(1 - \alpha)$ may not be possible. Indeed, in continuous settings we would like credible sets that do have exactly the right coverage, in order to minimize their size and thus obtain a more precise estimate. More generally, a technique for doing this is given by the *highest posterior density,* or HPD, credible set, defined as the set

$$C = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : p(\boldsymbol{\theta}|\mathbf{y}) \geq k(\alpha)\} ,$$

where $k(\alpha)$ is the largest constant satisfying

$$P(C|\mathbf{y}) \geq 1 - \alpha .$$

A good way to visualize such a set C is provided by Figure 2.3, which illustrates obtaining an HPD credible interval for θ assumed to follow a $G(2, 1)$ posterior distribution. Here, drawing a horizontal line across the graph at $p(\theta|\mathbf{y}) = 0.1$ results in a 87% HPD interval of (0.12, 3.59); see the solid line segments in the figure. That is, $P(0.12 < \theta < 3.59|\mathbf{y}) = .87$ when $\theta|\mathbf{y} \sim G(2, 1)$. We could obtain a 90% or 95% HPD interval simply by "pushing down" the horizontal line until the corresponding values on the θ-axis trapped the appropriate, larger probability.

Such a credible set is very intuitively appealing since it groups together the "most likely" θ values and is always computable from the posterior density, unlike classical CIs which frequently rely on "tricks" such as pivotal quantities. However, the graphical "line pushing" mentioned in the previous paragraph translates into iteratively solving the nonlinear equation

$$P(C(k(\alpha))|\mathbf{y}) = 1 - \alpha ,$$

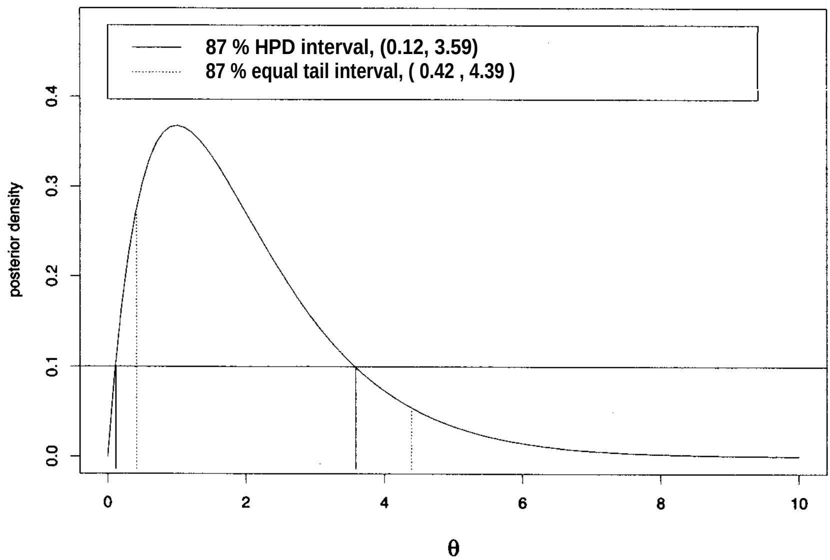


Figure 2.3 *HPD and equal-tail credible intervals arising from a Gamma(2,1) posterior using* $k(\alpha) = 0.1$.

in order to find C for a prespecified α level. This calculation is such a nuisance that we often instead simply take the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$ as our $100 \times (1 - \alpha)\%$ credible set for θ. This equal tail credible set will be equal to the HPD credible set if the posterior is symmetric and unimodal, but will be a bit wider otherwise.

In the context of our Figure 2.3 example, we obtain an 87% equal tail credible set of (0.42, 4.39); see the dashed line segments in the figure. By construction, we have

$$\begin{aligned} P(0.42 < \theta < 4.39|\mathbf{y}) &= 1 - P(\theta < 0.42|\mathbf{y}) - P(\theta > 4.39|\mathbf{y}) \\ &= 1 - \alpha/2 - \alpha/2 \\ &= 1 - \alpha \ = \ 0.87 \ . \end{aligned}$$

Due to the asymmetry of our assumed posterior distribution, this interval is 0.5 units wider than the corresponding HPD interval, and includes some larger θ values that are "less likely" while omitting some smaller θ values that are "more likely". Still, this interval is trivially computed for any a using one of several statistical software packages and seems to satisfactorily capture the "middle" of the posterior distribution. Moreover, this interval is invariant to 1-1 transformations of the parameter; for example, the logs of the interval endpoints above provide an 87% equal tail interval for $\eta \equiv \log \theta$.

For a typical Bayesian data analysis, we might summarize our findings by reporting (a) the posterior mean and mode, (b) several important posterior percentiles (corresponding to, say, probability levels .025, .25, .50, .75, and .975), (c) a plot of the posterior itself if it is multimodal, highly skewed, or otherwise badly behaved, and possibly (d) posterior probabilities of the form $P(\theta > c|\mathbf{y})$ or $P(\theta < c|\mathbf{y})$, where c is some important reference point (e.g., 0) that arises from the context of the problem. We defer an illustration of these principles until after introducing the basic concepts of Bayesian hypothesis testing.

2.3.3 Hypothesis testing and Bayes factors

The comparison of predictions made by alternative scientific explanations is, of course, a mainstay of current statistical practice. The traditional approach is to use the- ideas of Fisher, Neyman, and Pearson, wherein one states a primary, or *null,* hypothesis H_0, and an *alternative* hypothesis H_a. After determining an appropriate test statistic $T(\mathbf{Y})$, one then computes the *observed significance,* or *p-value*, of the test as

$$p\text{-value} = P\{T(\mathbf{Y}) \text{ more "extreme" than } T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\} , \tag{2.17}$$

where "extremeness" is in the direction of the alternative hypothesis. If the p-value is less than some prespecified Type I error rate, H_0 is rejected; otherwise, it is not.

While classical hypothesis testing has a long and celebrated history in the statistical literature and continues to be the overwhelming favorite of practitioners, several fairly substantial criticisms of it may be made. First, the approach can be applied straightforwardly only when the two hypotheses in question are *nested,* one within the other. That is, H_0 must normally be a simplification of H_a (say, by setting one of the model parameters in H_a equal to a constant - usually zero). But many practical hypothesis testing problems involve a choice between two (or more) models that are not nested (e.g., choosing between quadratic and exponential growth models, or between exponential and lognormal error distributions).

A second difficulty is that tests of this type can only offer evidence *against* the null hypothesis. A small p-value indicates that the larger, alternative model has significantly more explanatory power. However, a large p-value does not suggest that the two models are equivalent, but only that we lack evidence that they are not. This difficulty is often swept under the rug in introductory statistics courses, the technically correct phrase "fail to reject [the null hypothesis]" being replaced by "accept."

Third, the p-value itself offers no direct interpretation as a "weight of evidence," but only as a long-term probability (in a hypothetical repetition of the same experiment) of obtaining data at least as unusual as what was actually observed. Unfortunately, the fact that small p-values imply rejection of H_0 causes many consumers of statistical analyses to assume that the p-value is "the probability that H_0 is true," even though (2.17) shows that it is nothing of the sort.

A final, somewhat more philosophical criticism of p-values is that they depend not only on the observed data, but also the total sampling probability of certain *unobserved* datapoints; namely, the "more extreme" $T(\mathbf{Y})$ values in equation (2.17). As a result, two experiments with identical likelihoods could result in different p-values if the two experiments were *designed* differently. As mentioned previously in conjunction with Example 1.2, this fact violates a proposition known as the *Likelihood Principle* (Birnbaum, 1962), which can be stated briefly as follows:

> **The Likelihood Principle:** In making inferences or decisions about θ after $\mathbf{y}$ is observed, all relevant experimental information is contained in the likelihood function for the observed $\mathbf{y}$.

By taking into account not only the observed data y, but also the unobserved but more extreme values of $\mathbf{Y}$ classical hypothesis testing violates the Likelihood Principle.

This final criticism, while theoretically based, has practical import as well. For instance, it is not uncommon for unforeseen circumstances to arise during the course of an experiment, resulting in data that are not exactly what was intended in the experimental design. Technically speaking, in this case a frequentist analysis of the data is not possible, since p-values are only computable for data arising from the original design. Problems like this are especially common in the analysis of clinical trials data, where interim analyses and unexpected drug toxicities can wreak havoc with the trial's design. Berger and Berry (1988) provide a lucid and nontechnical discussion of these issues, including a clever example showing that a classically significant trial finding can be negated merely by *contemplating* additional trial stages that did not occur.

By contrast, the Bayesian approach to hypothesis testing, due primarily to Jeffreys (1961), is much simpler and more sensible in principle, and avoids all four of the aforementioned difficulties with the traditional approach. As in the case of interval estimation, however, it requires the notion of subjective probability to facilitate its probability calculus. Put succinctly, based on the data that each of the hypotheses is supposed to predict, one applies Bayes' Theorem and computes the posterior probability that the first hypothesis is correct. There is no limit on the number of hypotheses that may be simultaneously considered, nor do any need to be nested within any of the others. As such, in what follows, we switch notation from "hypotheses" H_0 and H_a to "models" $M_i,\ i = 1, \ldots, m$.

To lay out the mechanics more specifically, suppose we have two candidate parametric models M_1 and M_2 for data $\mathbf{Y}$ and the two models have respective parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Under prior densities $\pi_i(\boldsymbol{\theta}_i)$, $i = 1, 2$, the marginal distributions of $\mathbf{Y}$ are found by integrating out the parameters,

$$p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \ , \ i = 1, 2 \ . \tag{2.18}$$

Bayes' Theorem (2.1) may then be applied to obtain the posterior probabilities $P(M_1|\mathbf{y})$ and $P(M_2|\mathbf{y}) = 1 - P(M_1|\mathbf{y})$ for the two models. The quantity commonly used to summarize these results is the *Bayes factor, BF,* which is the ratio of the posterior odds of M_1 to the prior odds of M_1, given by Bayes' Theorem as

$$\begin{aligned} BF &= \frac{P(M_1|\mathbf{y})/P(M_2|\mathbf{y})}{P(M_1)/P(M_2)} & (2.19) \\ &= \frac{\left[\frac{p(\mathbf{y}|M_1)P(M_1)}{p(\mathbf{y})}\right] / \left[\frac{p(\mathbf{y}|M_2)P(M_2)}{p(\mathbf{y})}\right]}{P(M_1)/P(M_2)} \\ &= \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \ , & (2.20) \end{aligned}$$

the ratio of the observed marginal densities for the two models. Assuming the two models are *a priori* equally probable (i.e., $P(M_1) = P(M_2) = 0.5$), we have that $BF = P(M_1|\mathbf{y})/P(M_2|\mathbf{y})$, the posterior odds of M_1.

Consider the case where both models share the same parametrization (i.e. $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}$), and both hypotheses are simple (i.e., $M_1 : \boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}$ and $M_2 : \boldsymbol{\theta} = \boldsymbol{\theta}^{(2)}$). Then $\pi_i(\boldsymbol{\theta})$ consists of a point mass at $\boldsymbol{\theta}^{(i)}$ for $i = 1, 2$, and so from (2.18) and (2.20) we have

$$BF = \frac{f(\mathbf{y}|\boldsymbol{\theta}^{(1)})}{f(\mathbf{y}|\boldsymbol{\theta}^{(2)})} \ ,$$

which is nothing but the likelihood ratio between the two models. Hence, in the simple-versus-simple setting, the Bayes factor is precisely the odds in favor of M_1 over M_2 *given solely by the data.* For more general hypotheses, this same "evidence given by the data" interpretation of *BF is* often used, though Lavine and Schervish (1999) show that a more accurate interpretation is that *BF* captures the *change* in the odds in favor of model 1 as we move from prior to posterior. In any event, in such cases, *BF* does depend on the prior densities for the θ_i, which must be specified either as convenient conjugate forms or by more careful elicitation methods. In this case, it becomes natural to ask whether "shortcut" methods exist that provide a rough measure of the evidence in favor of one model over another without reference to any prior distributions.

One answer to this question is given by Schwarz (1978), who showed that for large sample sizes n, an approximation to -2 log BF *is* given by

$$\Delta BIC = W - (p_2 - p_1)\log n \ , \tag{2.21}$$

where p_i is the number of parameters in model M_i, i = 1, 2, and

$$W = -2\log\left[\frac{\sup_{M_1} f(\mathbf{y}|\boldsymbol{\theta})}{\sup_{M_2} f(\mathbf{y}|\boldsymbol{\theta})}\right] \ ,$$

the usual likelihood ratio test statistic. BIC stands for *Bayesian Information Criterion* (though it is also known as the *Schwarz Criterion),* and Δ denotes the change from Model 1 to Model 2. The second term in ΔBIC acts as a penalty term which corrects for differences in size between the models (to see this, think of M_2 as the "full" model and M_1 as the "reduced" model). Thus $\exp(-\frac{1}{2}\Delta BIC)$ provides a rough approximation to the Bayes factor which is independent of the priors on the θ_i.

There are many other penalized likelihood ratio model choice criteria, the most famous of which is the *Akaike Information Criterion,*

$$\Delta AIC = W - 2(p_2 - p_1) \ . \tag{2.22}$$

This criterion is derived from frequentist asymptotic efficiency considerations (where both n and the p_i approach infinity). This statistic is fairly popular in practice, but approaches -2 log BF asymptotically only if the information in the prior increases at the same rate. as the information in the likelihood, an unrealistic assumption in practice. Practitioners have also noticed that AIC tends to "keep too many terms in the model," as would be expected from its smaller penalty term.

While Bayes factors were viewed for many years as the only correct way to carry out Bayesian model comparison, they have come under increasing criticism of late. The most serious difficulty in using them is quite easy to see: if $\pi_i(\boldsymbol{\theta}_i)$ is improper (as is typically the case when using noninformative priors), then $p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$ necessarily is as well, and *so* BF as given in (2.20) is not well-defined. Several authors have attempted to modify the definition of BF to repair this deficiency. One approach is to divide the data into two parts, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ and use the first portion to compute model-specific posterior densities $p(\boldsymbol{\theta}_i|\mathbf{y}_1)$, i = 1, 2. Using these two posteriors as the priors in equation (2.20), the remaining data $\mathbf{y}_2$ then produces the *partial Bayes factor,*

$$BF(\mathbf{y}_2|\mathbf{y}_1) = \frac{p(\mathbf{y}_2|\mathbf{y}_1, M_1)}{p(\mathbf{y}_2|\mathbf{y}_1, M_2)} \ .$$

While this solves the nonexistence problem in the case of improper priors, there remains the issue of how many datapoints to allocate to the "training sample" $\mathbf{y}_1$ and how to select them. Berger and Pericchi (1996) propose using an arithmetic or geometric mean of the partial Bayes factors obtained using all possible minimal training samples (i.e., where n_1 is taken to be the smallest sample size leading to proper posteriors $p(\boldsymbol{\theta}_i|\mathbf{y}_1)$). Since the result corresponds to an actual Bayes factor under a reasonable intrinsic prior, they refer to it as an *intrinsic Bayes factor.* O'Hagan (1995) instead recommends a *fractional Bayes factor,* defined as

$$BF_b(\mathbf{y}) = \frac{p(\mathbf{y}, b|M_1)}{p(\mathbf{y}, b|M_2)} ,$$

where $b = n_1/n$, n_1 is again a "minimal" sample size, and

$$p(\mathbf{y}, b|M_i) = \frac{\int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)^b\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i} .$$

O'Hagan (1995, 1997) also suggests other, somewhat larger values for b, such as $b = \max(n_1, \log n)/n$ and $b = \max(n_1, \sqrt{n})/n$, on the grounds that this should reduce sensitivity to the prior.

Note that while the intrinsic Bayes factor uses part of the data to rectify the impropriety in the priors, the fractional Bayes factor instead uses a fraction b of the *likelihood function* for this purpose. O'Hagan (1997) offers asymptotic justifications for $BF_b(\mathbf{y})$ as well as comparison with various forms of the intrinsic *BF;* see also DeSantis and Spezzaferri (1999) for further comparisons, finite-sample motivations for using the fractional *BF,* and open research problems in this area.

2.3.4 Example: Consumer preference data

We summarize our presentation on Bayesian inference with a simple example, designed to illustrate key concepts in the absence of any serious computational challenges. Suppose sixteen consumers have been recruited by a fast food chain to compare two types of ground beef patty on the basis of flavor. All of the patties to be evaluated have been kept frozen for eight months, but one set of 16 has been stored in a high-quality freezer that maintains a temperature that is consistently within one degree of 0 degrees Fahrenheit. The other set of 16 patties has been stored in a freezer wherein the temperature varies anywhere between 0 and 15 degrees Fahrenheit. The food chain executives are interested in whether storage in the higher-quality freezer translates into a substantial improvement in taste (as judged by the consumers), thus justifying the extra effort and cost associated with equipping all of their stores with these freezers.

In a test kitchen, the patties are defrosted and prepared by a single chef. To guard against any possible order bias (e.g., it might be that a consumer will always prefer the first patty he or she tastes), each consumer is served the patties in a random order, as determined by the chef (who has been equipped with a fair coin for this purpose). Also, the study is *double-blind:* neither the consumers nor the servers know which patty is which; only the chef (who is not present at serving time) knows and records this

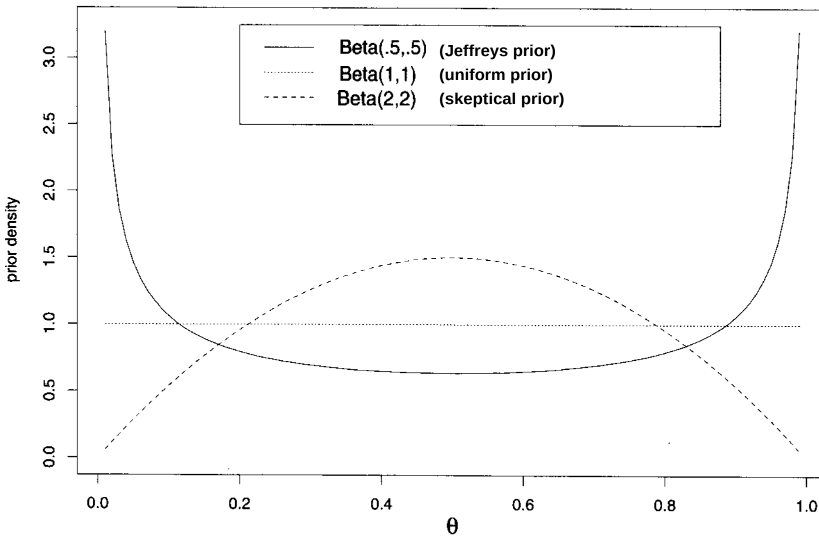


Figure 2.4 *Prior distributions used in analyzing the consumer preference data.*

information. This is also a measure to guard against possible bias on the part of the consumers or the servers (e.g., it might be that the server talks or acts differently when serving the cheaper patty). After allowing sufficient time for the consumers to make up their minds, the result is that 13 of the 16 consumers state a preference for the more expensive patty.

In order to analyze this data, we need a statistical model. In Bayesian analysis, this consists of two things: a likelihood and a prior distribution. Suppose we let θ be the probability that a consumer prefers the more expensive patty. Let $Y_i = 1$ if consumer i states a preference for the more expensive patty, and $Y_i = 0$ otherwise. If we assume that the consumers are independent and that θ is indeed constant over the consumers, then their decisions form a sequence of Bernoulli trials. Defining $X = \sum_{i=1}^{16} Y_i$, we have

$$X|\theta \sim Binomial(16, \theta) \; .$$

That is, the sampling density for x is $f(x|\theta) = \binom{16}{x}\theta^x(1-\theta)^{16-x}$.

Turning to the selection of a prior distribution, looking at the likelihood as a function of θ, we see that the beta distribution offers a conjugate family, since its density function is given by

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} .$$

Before adopting this family, however, we must ask ourselves if it is sufficiently broad to include our true prior feelings in this problem. Fortunately for us, the answer in this case seems to be yes, as evidenced by Figure 2.4. This figure shows three different members of the beta family, all of which might be appropriate. The solid line corresponds to the $Beta(.5, .5)$ distribution, which is the Jeffreys prior for this problem (see Section 2.6). We know this prior is noninformative in a transformation-invariate sense, but it seems to provide extra weight to extreme values of θ. The dotted line displays what we might more naturally think of as a noninformative prior, namely, a flat (uniform) prior, obtained by setting $\alpha = \beta = 1$. Finally, the dashed line follows a $Beta(2, 2)$ prior, which we might interpret as reflecting mild skepticism of an improvement in taste by the more expensive patty (i.e., it favors values of θ near .5). The beta family could also provide priors that are not symmetric about .5 (by taking $\alpha \neq \beta$), if the fast food chain executives had strong feelings or the results of previous studies to support such a conclusion. However, we shall restrict our attention to the three choices in Figure 2.4, since they seem to cover an adequate range of "minimally informative" priors which would not be controversial regardless of who is the ultimate decision-maker (e.g., the corporation's board of directors).

Thanks to the conjugacy of our prior family with the likelihood, the posterior distribution for θ is easily calculated as

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)\pi(\theta) \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{16-x+\beta-1} \\ &\propto Beta(x+\alpha, 16-x+\beta) . \end{aligned}$$

The three beta posteriors corresponding to our three candidate priors are plotted in Figure 2.5, with corresponding posterior quantile information summarized in Table 2.1. The three posteriors are ordered in the way we would expect, i.e., the evidence that θ is large decreases as the prior becomes more "skeptical" (concentrated near 1/2). Still, the differences among them are fairly minor, suggesting that despite the small sample size (n = 16), the results are robust to subtle changes in the prior. The posterior medians (which we might use as a point estimate of θ are all near 0.8; posterior modes (which would not account for the left skew in the distributions) would be even larger. In particular, the posterior mode under the flat prior is 13/16 = 0.8125, the maximum likelihood estimate of θ. Also note that all three posteriors produce 95% equal-tail credible intervals that exclude

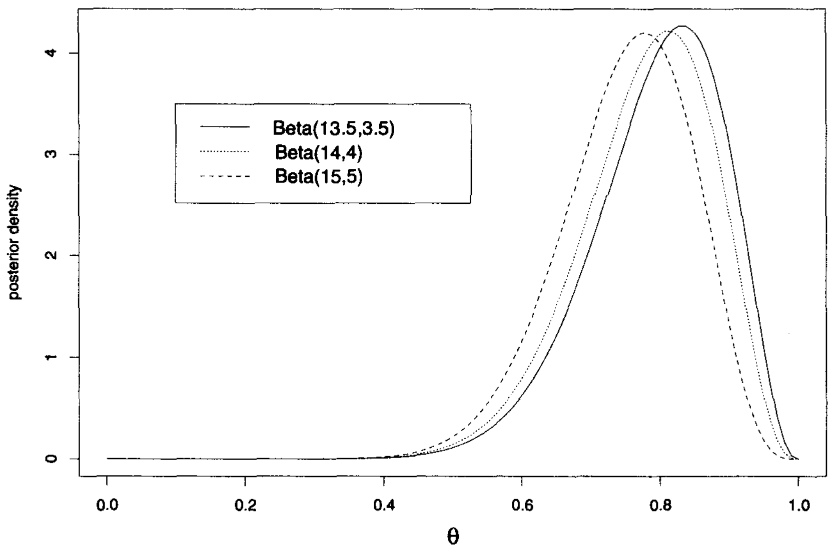


Figure 2.5 *Posterior distributions arising from the three prior distributions chosen for the consumer preference data.*

0.5, implying that we can be quite confident that the more expensive patty offers an improvement in taste.

To return to the executives' question concerning a *substantial* improvement in taste, we must first define "substantial." We (somewhat arbitrarily) select 0.6 as the critical value which θ must exceed in order for the improvement to be regarded as "substantial." Given this cutoff value, it is natural to compare the hypotheses $M_1 : \theta \geq 0.6$ and $M_2 : \theta < 0.6$. Using the uniform prior, the Bayes factor in favor of M_1 is easily computed using

Prior distribution	Posterior quantile .025	.500	.975	$P(\theta > .6 \mid x)$
$Beta(.5, .5)$	0.579	0.806	0.944	0.964
$Beta(1, 1)$	0.566	0.788	0.932	0.954
$Beta(2, 2)$	0.544	0.758	0.909	0.930

Table 2.1 *Posterior summaries, consumer preference data.*

equation (2.19) and Table 2.1 as

$$BF = \frac{0.954/0.046}{0.4/0.6} = 31.1 \ ,$$

fairly strong evidence in favor of a substantial improvement in taste. Jeffreys suggested judging BF in units of 0.5 on the common-log-odds scale. Since $\log_{10} 31.1 \approx 1.5$, this rule also implies a strong preference for M_1.

2.4 Model assessment

2.4.1 Diagnostic measures

Having successfully fit a model to a given dataset, the statistician must be concerned with whether or not the fit is adequate and the assumptions made by the model are justified. For example, in standard linear regression, the assumptions of normality, independence, linearity, and homogeneity of variance must all be investigated. Several authors have suggested using the marginal distribution of the data, $m(\mathbf{y})$, in this regard. Observed y_i values for which $m(y_i)$ is small are "unlikely" and, therefore, may be considered outliers under the assumed model. Too many small values of $m(y_i)$ suggest the model itself is inadequate and should be modified or expanded.

A problem with this approach is the difficulty in defining how small is "small," and how many outliers are "too many." In addition, we have the problem of the possible impropriety of $m(\mathbf{y})$ under noninformative priors, already mentioned in the context of Bayes factors. As such, we might instead work with *predictive* distributions, since they will be proper whenever the posterior is. Specifically, suppose we fit our Bayesian model using a *fitting sample* of data $\mathbf{z} = (z_1, \ldots, z_m)'$ and wish to check it using an independent *validation sample* $\mathbf{y} = (y_1, \ldots, y_n)'$. By analogy with classical model checking, we could begin by analyzing the collection of *residuals,* the difference between the observed and fitted values for each point in the validation sample. In this context, we define a Bayesian residual as

$$r_i = y_i - E(Y_i|\mathbf{z}), \ i = 1, \ldots, n. \qquad (2.23)$$

Plotting these residuals versus fitted values might reveal failure in a normality or homogeneity of variance assumption; plotting them versus time could reveal a failure of independence. Summing their squares or absolute values could provide an overall measure of fit, but, to eliminate the dependence on the scale of the data, we might first *standardize* the residuals, computing

$$d_i = \frac{y_i - E(Y_i|\mathbf{z})}{\sqrt{Var(Y_i|\mathbf{z})}}, \ i = 1, \ldots, n. \qquad (2.24)$$

The above discussion assumes the existence of two independent data samples, which may well be unavailable in many problems. Of course we could simply agree to always reserve some portion (say, 20-30%) of our data at the outset for subsequent model validation, but we may be loathe to do this when data are scarce. As such, Gelfand, Dey, and Chang (1992) suggest a *cross- validatory* (or "leave one out") approach, wherein the fitted value for y_i is computed conditional on all the data *except* y_i, namely, $\mathbf{y}_{(i)} \equiv (y_1, \ldots y_{i-1}, y_{i+1}, \ldots, y_n)'$. That is, we replace (2.23) by

$$r_i' = y_i - E(Y_i|\mathbf{y}_{(i)}) \ ,$$

and (2.24) by

$$d_i' = \frac{y_i - E(Y_i|\mathbf{y}_{(i)})}{\sqrt{Var(Y_i|\mathbf{y}_{(i)})}} \ .$$

Gelfand et al. (1992) provide a collection of alternate definitions for di, with associated theoretical support presented in Gelfand and Dey (1994).

Note that in this cross-validatory approach we compute the posterior mean and variance with respect to the *conditional predictive distribution,*

$$f(y_i|\mathbf{y}_{(i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(i)})} = \int f(y_i|\boldsymbol{\theta}, \mathbf{y}_{(i)}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \ , \qquad (2.25)$$

which gives the likelihood of each point given the remainder of the data. The actual values of $f(y_i|\mathbf{y}_{(i)})$ (sometimes referred to as the *conditional predictive ordinate,* or CPO) can be plotted versus i as an outlier diagnostic, since data values having low CPO are poorly fit by the model. Notice that, unlike the marginal distribution, $f(y_i|\mathbf{y}_{(i)})$ will be proper if $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ is. In addition, the collection of conditional predictive densities $\{f(y_i|\mathbf{y}_{(i)}),\ i = 1, \ldots, n\}$ is equivalent to $m(\mathbf{y})$ when both exist (Besag, 1974), encouraging the use of the former even when the latter is undefined.

An alternative to this cross-validatory approach is due to Rubin (1984), who proposes instead *posterior predictive* model checks. The idea here is to construct test statistics or other "discrepancy measures" $D(\mathbf{y})$ that attempt to measure departures of the observed data from the assumed model (likelihood and prior distribution). For example, suppose we have fit a normal distribution to a sample of univariate data, and wish to investigate the model's fit in the lower tail. We might compare the observed value of the discrepancy measure

$$D(\mathbf{y}) = y_{min}$$

with its posterior predictive distribution, $p(D(\mathbf{y}^*)|\mathbf{y})$, where $\mathbf{y}^*$ denotes a hypothetical future value of y. If the observed value is extreme relative to this reference distribution, doubt is cast on some aspect of the model. Gelman et al. (1995) stress that the choice of $D(\mathbf{y})$ can (and should) be tuned to the aspect of the model whose fit is in question. However, $D(\mathbf{y})$ should not focus on aspects which are parametrized by the model (e.g., taking $D(\mathbf{y})$ equal to the sample mean or variance in the above example), since these would be fit automatically in the posterior distribution.

In order to be computable in the classical framework, test statistics must be functions of the observed data alone. But, as pointed out by Gelman, Meng, and Stern (1996), basing Bayesian model checking on the posterior predictive distribution allows generalized test statistics $D(\mathbf{y}, \boldsymbol{\theta})$ that depend on the parameters as well as the data. For example, as an omnibus goodness-of-fit measure, Gelman et al. (1995, p.172) recommend

$$D(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{[y_i - E(Y_i|\boldsymbol{\theta})]^2}{Var(Y_i|\boldsymbol{\theta})} . \tag{2.26}$$

With $\boldsymbol{\theta}$ varying according to its posterior distribution, we would now compare the distribution of $D(\mathbf{y}, \boldsymbol{\theta})$ for the observed y with that of $D(\mathbf{y}^*, \boldsymbol{\theta})$ for a future observation $\mathbf{y}^*$. A convenient summary measure of the extremeness of the former with respect to the latter is the tail area,

$$\begin{aligned} p_D &\equiv P[D(\mathbf{y}^*, \boldsymbol{\theta}) > D(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}] \\ &= \int P[D(\mathbf{y}^*, \boldsymbol{\theta}) > D(\mathbf{y}, \boldsymbol{\theta})|\theta]\, p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} . \end{aligned} \tag{2.27}$$

In the case where $D(\mathbf{y}^*, \boldsymbol{\theta})$'s distribution is free of $\boldsymbol{\theta}$ (in frequentist parlance, when $D(\mathbf{y}^*, \boldsymbol{\theta})$ is a *pivotal quantity*), p_D is exactly equal to the frequentist p-value, or the probability of seeing a statistic as extreme as the one actually observed. As such, (2.27) is sometimes referred to as the *Bayesian p-value.*

Given our earlier rejection of the p-value (in Example 1.2 and elsewhere), primarily for its failure to obey the Likelihood Principle, our current interest in a quantity that may be identical to it may seem incongruous. But as stressed by Meng (1994), it is important to remember that we are not advocating the use of (2.27) in model choice: p_D is *not* the probability that the model is "correct," and Bayesian p-values should *not* be compared across models. Rather, they serve only as measures of discrepancy between the assumed model and the observed data, and hence provide information concerning model adequacy. Even on this point there is controversy, since, strictly speaking, this model checking strategy uses the data twice: once to compute the observed statistic $D(\mathbf{y}, \boldsymbol{\theta})$, and again to obtain the posterior predictive reference distribution. It is for this reason that, when sufficient data exist, we prefer the approach that clearly separates the fitting sample $\mathbf{z}$ from the validation sample y. Then we could find

$$\begin{aligned} p'_D &\equiv P[D(\mathbf{y}^*, \boldsymbol{\theta}) > D(\mathbf{y}, \boldsymbol{\theta})|\mathbf{z}] \\ &= \int P[D(\mathbf{y}^*, \boldsymbol{\theta}) > D(\mathbf{y}, \boldsymbol{\theta})|\theta]\, p(\boldsymbol{\theta}|\mathbf{z}) d\boldsymbol{\theta} . \end{aligned} \tag{2.28}$$

Still, (2.27) could be viewed merely as an approximation of the type found in empirical Bayes analyses; see Chapter 3. Moreover, the generalized form in (2.27) accommodates nuisance parameters and provides a Bayesian probability interpretation in terms of future datapoints, thus somewhat rehabilitating the p-value.

2.4.2 *Model averaging*

The discussion so far in this section suggests that the line between model adequacy and model choice is often a blurry one. Indeed, some authors (most notably, Gelfand, Dey, and Chang, 1992) have suggested cross-validatory model adequacy checks as an informal alternative to the more formal approach to model choice offered by Bayes factors and their variants. We defer details until Subsection 6.5.

In some problem settings there may be underlying scientific reasons why one model or another should be preferred (for example, a spatial distribution of observed disease cases, reflecting uncertainty as to the precise number of disease "clusters" or sources of contamination). But in other cases, the process of model selection may be somewhat arbitrary in that a number of models may fit equally well and simultaneously provide plausible explanations of the data (for example, a large linear regression problem with 20 possible predictor variables). In such settings, choosing a single model as the "correct" one may actually be suboptimal, since further inference conditional on the chosen model would of necessity ignore the uncertainty in the model selection process itself. This in turn would lead to overconfidence in our estimates of the parameters of interest.

In principle, the Bayesian approach can handle this difficulty simply by replacing model choice with *model averaging.* Suppose $(M_1, \ldots, M_m)$ again index our collection of candidate models and γ is a quantity of interest, assumed to be well-defined for every model. Given a set of prior model probabilities $\{p(M_1), \ldots, p(M_m)\}$, the posterior distribution of γ is

$$p(\gamma|\mathbf{y}) = \sum_{i=1}^{m} p(\gamma|M_i, \mathbf{y})p(M_i|\mathbf{y}) \ , \tag{2.29}$$

where $p(\gamma|M_i, \mathbf{y})$ is the posterior for γ under the i^{th} model, and $p(M_i|\mathbf{y})$ is the posterior probability of this model, computable as

$$p(M_i|\mathbf{y}) = \frac{p(\mathbf{y}|M_i)p(M_i)}{\sum_{j=1}^{m} p(\mathbf{y}|M_j)p(M_j)} \ ,$$

where $p(\mathbf{y}|M_i)$ is the marginal distribution of the data under the i^{th} model, given previously in equation (2.18). Suppose we rank models based on the log of their conditional predictive ordinate evaluated at γ, namely, $\log p(\gamma|M_i, \mathbf{y})$. Using the nonnegativity of the Kullback-Leibler divergence, Madigan and Raftery (1994) show that for each model $i = 1, \ldots, m$,

$$E\left[\log\left(\sum_{j=1}^{m} p(\gamma|M_j, \mathbf{y})p(M_j|\mathbf{y})\right)\right] \geq E\left[\log p(\gamma|M_i, \mathbf{y})\right] ,$$

where the expectations are with respect to the joint posterior distribution of γ and the M_is. That is, averaging over all the models results in a better log predictive score than using any one of the models individually.

Despite this advantage, there are several difficulties in implementing a solution via equation (2.29). All of them basically stem from the fact that the number of potential models m is often extremely large. For example, in our 20-predictor regression setting, $m = 2^{20}$ (since each variable can be in or out of the model), and this does not count variants of each model arising from outlier deletion or variable transformation. Specifying the prior probabilities for this number of models is probably infeasible unless some fairly automatic rule is implemented (say, $p(M_i) = 1/m$ for all i). In addition, direct computation of the sum in (2.29) may not be possible, forcing us to look for a smaller summation with virtually the same result. One possibility (also due to Madigan and Raftery, 1994) is to limit the class of models to those which meet some minimum level of posterior support, e.g., the class

$$\mathcal{A}' = \left\{M_i : \frac{\max_j p(M_j|\mathbf{y})}{p(M_i|\mathbf{y})} \leq C\right\} ,$$

say for $C = 20$. They further eliminate the set $\mathcal{B}$ of models that have less posterior support than their own submodels, and then perform the summation in (2.29) over the set $\mathcal{A} = \mathcal{A}' \backslash \mathcal{B}$. Madigan and Raftery (1994) give a greedy-search algorithm for locating the elements of $\mathcal{A}$, and offer several examples to show that their procedure loses little of the predictive power available by using the entire model collection.

An alternative would be to search the entire sample space using Markov chain Monte Carlo methods, locating the high-probability models and including them in the summation. Since we have not yet presented these methods, it is difficult to say much more about them at this time. Indeed, many of the predictive quantities described in this section involve computation of high-dimensional integrals that are routinely evaluated via Monte Carlo methods. As such, we defer further discussion of these issues until Chapter 6, after we have developed the computational machinery to enable routine calculation of the residuals and predictive summaries in Chapter 5. We do however point out that all of the model averaging approaches discussed so far assume that the search may be done in "model only" space (i.e., the parameters $\boldsymbol{\theta}$ can be integrated out analytically prior to beginning the algorithm). For searches in "model-parameter" space, matters are more complicated; algorithms for this situation are provided by George and Mc-

Culloch (1993), Carlin and Chib (1995), Chib (1995), Green (1995), Phillips and Smith (1996), Dellaportas et al. (2000), and Godsill (2000).

2.5 Nonparametric methods

The previous section describes how we might mix over various models M_i in obtaining a posterior or predictive distribution that more accurately reflects the underlying uncertainty in the precise choice of model. In a similar spirit, it might be that no single parametric model $f(y|\theta)$ could fully capture the variability inherent in y's conditional distribution, suggesting that we average over a collection of such densities. For example, for the sake of convenience we might average over normal densities,

$$f(y|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{q}) = \sum_{k=1}^{K} q_k \, \phi(y|\mu_k, \sigma_k^2) \ , \tag{2.30}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)$, $\mathbf{q} = (q_1, \ldots, q_K)$, and $\phi(.|.,.)$ denotes the normal density function. Equation (2.30) is usually referred to as a *mixture* density, which has mixing probabilities q_k satisfying $\sum_{k=1}^{K} q_k = 1$. Mixture models are attractive since they can accommodate an arbitrarily large range of model anomalies (multiple modes, heavy tails, and so on) simply by increasing the number of components in the mixture. They are also quite natural if the observed population is more realistically thought of as a combination of several distinct subpopulations (say, a collection of ethnic groups). An excellent reference on the Bayesian analysis of models of this type is the book by Titterington, Makov, and Smith (1985).

Unfortunately, inference using mixture distributions like (2.30) is often difficult. Since increasing one of the σ_k^2 and increasing the number of mixture components K can have a similar effect on the shape of the mixture, the parameters of such models are often not fully identified by the data. This forces the adoption of certain restrictions, such as requiring $\sigma_k^2 = \sigma^2$ for all k, or fixing the dimension K ahead of time. These restrictions in turn lead to difficulties in computing the relevant posterior and predictive densities; in this regard see West (1992) and Diebolt and Robert (1994) for a list of potential problem areas and associated remedies.

An appealing alternative to discrete finite mixtures like (2.30) would be *continuous* mixtures of the form

$$f(y|g) = \int f(y|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \ .$$

But this creates the problem of choosing an appropriate mixing distribution g. Of course we could select a member of some parametric family $g(\boldsymbol{\theta}|\eta)$, but the appropriate choice might not be apparent before seeing the data. As such, we might wish to specify g *nonparametrically*, in effect, allowing the data to suggest the appropriate mixing distribution.

The most popular method for doing this is via the *Dirichlet process* (DP) prior, introduced by Ferguson (1973). This prior is a distribution on the set of *all* probability distributions $\mathcal{P}$. Suppose we are given a base distribution G_0 and a scalar precision parameter ν. Then, for any finite partition of the parameter space $\{B_1, \ldots, B_m\}$ the DP prior is a random probability measure $G \in \mathcal{P}$ that assigns a Dirichlet distribution with m-dimensional parameter vector $(\nu G_0(B_1), \ldots, \nu G_0(B_m))$ to the vector of probabilities $(G(B_1), \ldots, G(B_m))$. (The Dirichlet distribution is a multivariate generalization of the beta distribution; see Appendix A.) A surprising consequence of this formulation is that the DP assigns probability 1 to the subset of *discrete* probability measures. We use the standard notation $G \sim DP(\nu G_0)$ to denote this Dirichlet process prior for G.

The use of the Dirichlet process in the context of mixture distributions essentially began with Antomak (1974). Here, we work with the set of probability distributions $\mathcal{F}$ for y having densities of the form

$$f(y|G) = \int f(y|\boldsymbol{\theta})dG(\boldsymbol{\theta}) ,$$

where $G \sim DP(\nu G_0)$. Hence $\mathcal{F}$ is a nonparametric mixture family, with $G(\boldsymbol{\theta})$ playing the role of the conditional distribution of $\boldsymbol{\theta}$ given G. Note that we may partition $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and DP mix on only the subvector $\boldsymbol{\theta}_1$, placing a traditional parametric prior on $\boldsymbol{\theta}_2$; many authors refer to such models as *semiparametric*.

In practice, we will usually have independent data values $y_i, i = 1,\ldots,n$. Inserting a latent parameter $\boldsymbol{\theta}_i$ for each y_i we obtain the hierarchical model

$$\begin{aligned} Y_i|\boldsymbol{\theta}_i &\overset{ind}{\sim} f_i(y_i|\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|G &\overset{iid}{\sim} G(\boldsymbol{\theta}) \\ G|G_0,\nu &\sim DP(\nu G_0) \end{aligned} \tag{2.31}$$

where here we DP mix over the entire $\boldsymbol{\theta}$ vector.

Computational difficulties precluded the widespread use of DP mixture models until very recently, when a series of papers (notably Escobar, 1994, and Escobar and West, 1992, 1995) showed how Markov chain Monte Carlo methods could be used to obtain the necessary posterior and predictive distributions. These methods avoid the infinite dimension of G by analytically marginalizing over this dimension in the model, obtaining the likelihood

$$\left[\prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}_i)\right] p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G_0, \nu) . \tag{2.32}$$

Note that this marginalization implies that the $\boldsymbol{\theta}_i$ are no longer conditionally independent as in (2.31). Specifying a prior distribution on ν and the parameters of the base distribution G_0 completes the Bayesian model specification. The basic computational algorithm requires that G_0 be conjugate with the f_i, but a recent extension due to MacEachern and Müller (1994) eliminates this restriction (though, of course, not without additional algorithmic complexity). Most recently, Mukhopadhyay and Gelfand (1997) combine DP mixtures with overdispersed exponential family models (Gelfand and Dalal, 1990), producing an extremely broad extension of the generalized linear model setting.

DP mixing is an extremely recent addition to the applied Bayesian toolkit, and many practical issues remain to be addressed. For example, note that large values of the DP precision parameter ν favor adoption of the base measure G_0 as the posterior for the $\boldsymbol{\theta}_i$, while small values will produce the empirical cdf of the data as this posterior. As such, the precise choice of the prior distribution for ν (and the sensitivity of the results to this decision) is still an area of substantial debate. Other unresolved issues include the proper selection of computational algorithm and associated convergence checks, and the suitability and feasibility of the models in real data settings. Still, the substantial current research activity in nonparametric Bayesian methods suggests a bright future in practice.

2.6 Exercises

1. Let $Y_1, Y_2 \ldots$ be a sequence of random variables (not necessarily independent) taking on the values 0 and 1. Let $S_n = Y_1 + \ldots Y_n$.

   (a) Represent $P(Y_{n+1} = 1|S_n = s)$ in terms of the joint distribution of the Y_i.

   (b) Further refine the representation when all permutations of the Y_i such that $S_n = s$ have equal probability (i.e., the Y_i are exchangeable).

   (c) What is the conditional probability if the Y_i are independent?

2. Confirm expression (2.2) for the conditional posterior distribution of θ given σ^2 in the conjugate normal model, and show that the precision in this posterior is the sum of the precisions in the likelihood and the normal prior, where *precision* is defined to be the reciprocal of the variance.

3. Confirm expression (2.9) for the conditional posterior distribution of given θ in the conjugate normal model. If your random number generator could produce only χ^2 random variates, how could you obtain samples from (2.9)?

4. Let θ be the height of the tallest individual currently residing in the city where you live. Elicit your own prior density for θ by

   (a) a histogram approach

   (b) matching an appropriate distributional form from Appendix A.

   What difficulties did you encounter under each approach?

5. For each of the following densities from Appendix A, provide a conjugate prior distribution for the unknown parameter(s), if one exists:

(a) $X \sim Bin(n, \theta)$, n known
(b) $X \sim NegBin(r, \theta)$, r known
(c) $\mathbf{X} \sim Mult(n, \boldsymbol{\theta})$, n known
(d) $X \sim G(\alpha, \beta)$, α known
(e) $X \sim G(\alpha, \beta)$, β known
(f) $\mathbf{X} \sim N_k(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ known
(g) $\mathbf{X} \sim N_k(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\boldsymbol{\theta}$ known

6. For the following densities from Appendix A, state whether each represents a location family, scale family, location-scale family, or none of these. For any of the first three situations, suggest a simple noninformative prior for the unknown parameter(s):

(a) $X \sim DE(1, \sigma^2)$
(b) $X \sim Unif(\theta - a, \theta + a)$
(c) $X \sim Logistic(\mu, 1)$
(d) $X \sim G(\alpha, \beta)$, α known
(e) $X \sim G(\alpha, \beta)$, β known

7. Let θ be a univariate parameter of interest, and let $\gamma = g(\theta)$ be a 1-1 transformation. Use (2.10) and (2.11) to show that (2.12) holds, i.e., that the Jeffreys prior is invariant under reparametrization. (*Hint:* What is the expectation of the so-called *score statistic*, $d \log f(\mathbf{x}|\theta)/d\theta$?)

8. Suppose that $f(x|\theta, \sigma)$ is normal with mean θ and standard deviation σ.

(a) Suppose that σ is known, and show that the Jeffreys prior (2.10) for θ is given by

$$p(\theta) = 1,\ \theta \in \Re\ .$$

(b) Next, suppose that θ is known, and show that the Jeffreys prior for σ is given by

$$p(\sigma) = \frac{1}{\sigma},\ \sigma > 0\ .$$

(c) Finally, assume that θ and σ are both unknown. and show that the bivariate Jeffreys prior (2.13) is given by

$$p(\theta, \sigma) = \frac{1}{\sigma^2},\ \theta \in \Re,\ \sigma > 0\ ,$$

a slightly different form than that obtained by simply multiplying the two individual Jeffreys priors obtained above.

9. Show that the Jeffreys prior based on the binomial likelihood $f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ is given by the $Beta(.5, .5)$ distribution.

10. Suppose that in the situation of Example *2.2* we adopt the noninformative prior $p(\theta, \sigma) = \frac{1}{\sigma}$, $\theta \in \Re$, $\sigma > 0$ (i.e., $p(\theta, \sigma^2) = \frac{1}{\sigma^2}$, $\theta \in \Re$, $\sigma^2 > 0$).

   (a) Show that the marginal posterior for $t = \sqrt{n}(\theta - \bar{y})/s$, where $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$, is proportional to

   $$[1 + t^2/(n-1)]^{-n/2} \; ,$$

   a Student's t distribution with *(n - 1)* degrees of freedom.

   (b) Show that the marginal posterior for σ^2 is proportional to

   $$(\sigma^2)^{-[(n-1)/2+1]} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2\right] \; ,$$

   an $IG\left(\frac{n-1}{2}, \frac{2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right)$ density.

11. Suppose that $Y|\theta \sim G(1, \theta)$ (i.e., the *exponential* distribution with mean θ), and that $\theta \sim IG(\alpha, \beta)$.

   (a) Find the posterior distribution of θ.

   (b) Find the posterior mean and variance of ^

   (c) Find the posterior mode of θ.

   (d) Write down two integral equations that could be solved to find the 95% equal-tail credible interval for θ.

12. Suppose $X_1, \ldots, X_n \overset{iid}{\sim}$ *NegBin*(r, θ), r known, and that θ is assigned a *Beta*(α, β) prior distribution.

   (a) Find the posterior distribution of θ.

   (b) Suppose further that r $= 5$, $n = 10$, and we observe $\sum_{i=1}^{n} x_i = 70$. Of the two hypotheses $H_1 : \theta \leq .5$ and $H_2 : \theta > .5$, which has greater posterior probability under the mildly informative $Beta(2, 2)$ prior?

   (c) What is the Bayes factor in favor of H_1? Does it suggest strong evidence in favor of this hypothesis?

13. Consider the two-stage binomial problem wherein $\theta \sim G$, and $X_i|\theta \overset{iid}{\sim}$ *Bernoulli*(θ), $i = 1, \ldots, n$.

   (a) Find the joint marginal distribution of $\mathbf{X}_n = (X_1, \ldots, X_n)$.

   (b) Are the components independent? In any case, describe features of the marginal distribution.

   (c) Let $S_n = X_1 + \cdots + X_n$. Under what conditions does S_n have a binomial distribution? Describe features of its distribution.

   (d) If $dG(\theta) = d\theta$, find the distribution of S_n and discuss its features.

14. For the Dirichlet process prior described in Section *2.5*, use the properties of the Dirichlet distribution to show that for any measurable set *B*,

(a) $E[G(B)] = G_0(B)$

(b) $Var[G(B)] = G_0(B)(1 - G_0(B))/(\nu + 1)$

What might be an appropriate choice of vague prior distribution on ν?

15. The data in Table 2.2 come from a 1980 court case (Reynolds v. Sheet Metal Workers) involving alleged applicant discrimination. Perform a Bayesian analysis of these data. Is there evidence of discrimination?

	Selected	Rejected	Total
Black	14	30	44
White	41	39	80
Total	55	69	124

Table 2.2 *Summary table from Reynolds v. Sheet Metal Workers*

16. The following are increased hours of sleep for 10 patients treated with soporific B compared with soporific A (Cushny and Peebles, presented by Fisher in *Statistical Methods for Research Workers*):

$$1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4$$

Perform a Bayesian analysis of this data and draw conclusions, assuming each component of the likelihood to be

(a) normal

(b) Student's t with 3 degrees of freedom (t_3)

(c) Cauchy (t_1)

(d) Bernoulli (i.e., positive with some probability θ), dealing with the 0 somehow (for example, by ignoring it).

Don't forget to comment on relevant issues in model assessment and model selection.

(*Hint:* Write a Fortran, C, ***Matlab***, or ***S-plus*** program to handle any necessary numerical integration and computation of marginal posterior probabilities. Subroutine libraries or statistical packages may be helpful for the integration or plotting of marginal posterior distributions.)

CHAPTER 3

The empirical Bayes approach

3.1 Introduction

As detailed in Chapter 2, in addition to the likelihood, Bayesian analysis depends on a prior distribution for the model parameters. This prior can be nonparametric or parametric, depending on unknown parameters which may in turn be drawn from some second-stage prior. This sequence of parameters and priors constitutes a hierarchical model. The hierarchy must stop at some point, with all remaining prior parameters assumed known. Rather than make this assumption, the empirical Bayes (EB) approach uses the observed data to estimate these final stage parameters (or to directly estimate the Bayes decision rule) and then proceeds as though the prior were known.

Though in principle EB can be implemented for hierarchical models having any number of levels, for simplicity we present the approach for the standard two-stage model. That is, we assume a likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ for the observed data y given a vector of unknown parameters $\boldsymbol{\theta}$ and a prior for $\boldsymbol{\theta}$ with cdf $G(\boldsymbol{\theta})$ and corresponding density or mass function $g(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of hyperparameters. With $\boldsymbol{\eta}$ known, the Bayesian uses Bayes' rule (2.1) to compute the posterior distribution,

$$p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\eta}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})}{m(\mathbf{y}|\boldsymbol{\eta})}, \tag{3.1}$$

where $m(\mathbf{y}|\boldsymbol{\eta})$ denotes the marginal distribution of y,

$$m(\mathbf{y}|\boldsymbol{\eta}) = \int f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta}, \tag{3.2}$$

or, in more general notation, $m_G(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta})dG(\boldsymbol{\theta})$.

If $\boldsymbol{\eta}$ is unknown, the fully Bayesian (Chapter 2) approach would adopt a hyperprior distribution, $h(\boldsymbol{\eta})$, and compute the posterior distribution as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})h(\boldsymbol{\eta})d\boldsymbol{\eta}}{\int\int f(\mathbf{y}|\mathbf{u})g(\mathbf{u}|\boldsymbol{\eta})h(\boldsymbol{\eta})d\mathbf{u}d\boldsymbol{\eta}} = \int p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\eta})h(\boldsymbol{\eta}|\mathbf{y})d\boldsymbol{\eta}. \tag{3.3}$$

The second representation shows that the posterior is a mixture of conditional posteriors (3.1) given a fixed $\boldsymbol{\eta}$, with mixing via the marginal posterior distribution of $\boldsymbol{\eta}$.

In empirical Bayes analysis, we instead use the marginal distribution (3.2) to estimate $\boldsymbol{\eta}$ by $\hat{\boldsymbol{\eta}} \equiv \hat{\boldsymbol{\eta}}(\mathbf{y})$ (for example, the marginal maximum likelihood estimator). Inference is then based on the *estimated posterior* distribution $p(\boldsymbol{\theta}|\mathbf{y}, \hat{\boldsymbol{\eta}})$. The EB approach thus essentially replaces the integration in the rightmost part of (3.3) by a maximization, a substantial computational simplification. The name "empirical Bayes" arises from the fact that we are using the data to estimate the hyperparameter $\boldsymbol{\eta}$. Adjustments to account for the uncertainty induced by estimating $\boldsymbol{\eta}$ may be required, especially to produce valid interval estimates (see Section 3.5).

The development in the preceding paragraph actually describes what Morris (1983a) refers to as the *parametric* EB (PEB) approach. That is, we assume the distribution $g(\boldsymbol{\theta}|\boldsymbol{\eta})$ takes a parametric form, so that choosing a (data-based) value for $\boldsymbol{\eta}$ is all that is required to completely specify the estimated posterior distribution. For more flexibility, one can implement a *nonparametric* EB (NPEB) approach, in which $G(\boldsymbol{\theta})$ has an unknown form. Pioneered and championed by Robbins (1955, 1983). and further generalized and modernized by Maritz and Lwin (1989, Section 3.4). van Houwelingen (1977) and others, this method first attempts to represent the posterior mean in terms of the marginal distribution, and then uses the data to estimate the Bayes rule directly. Recent advances substitute a nonparametric estimate of $G(\boldsymbol{\theta})$, generalizing the approach to a broader class of models and inferential goals. The nonparametric setup is discussed in the next section, with the remainder of the chapter primarily devoted to the parametric EB formulation.

3.2 Nonparametric EB (NPEB) point estimation

3.2.1 Compound sampling models

Since, for EB analysis, $G(\boldsymbol{\theta})$ is not completely known and at least some of its features must be estimated, repeated draws from G are required to obtain sufficient information. Consider then the *compound sampling model,*

$$\begin{aligned} Y_i \mid \theta_i \quad &\overset{ind}{\sim} \quad f_i(y_i \mid \theta_i)\,, \quad \text{and} \\ \theta_i \quad &\overset{iid}{\sim} \quad G(\cdot)\,, \end{aligned} \tag{3.4}$$

where $i = 1, \ldots, k$. Compound sampling arises in a wide variety of applications including multi-site clinical trials, estimation of disease rates in small geographic areas, longitudinal studies, laboratory assays, and meta-analysis. Model (3.4) can be extended by allowing regression and correlation structures in G, or allowing the Y_i to be correlated given the θ_i.

Suppose we seek point estimates for the θ_i, and that the prior cdf G has corresponding density function g. Writing $\mathbf{Y} = (Y_1, \ldots, Y_k)$ and $\boldsymbol{\theta} =$

$(\theta_1, \ldots, \theta_k)$, it is left as an exercise to show that the marginal density of the data $m_G(\mathbf{y})$ satisfies the equation

$$m_G(\mathbf{y}) = \prod_{i=1}^{k} m_G(y_i) \; , \qquad (3.5)$$

where $m_G(y_i) = \int f_i(y_i \mid \theta_i) dG(\theta_i)$. That is, the Y_i are marginally independent (and identically distributed as well if $f_i = f$ for all *i*).

3.2.2 *Simple NPEB (Robbins' method)*

Consider the basic compound model (3.4) where *G* is completely unspecified, and $Y_i|\theta_i$ is Poisson(θ_i). That is,

$$f(y_i \mid \theta_i) = \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!}, \; y_i = 0, 1, 2, \ldots.$$

As shown in Appendix B, the Bayes estimate of θ_i under squared error loss is the posterior mean,

$$\begin{aligned} \theta_i^B &= E(\theta_i \mid \mathbf{y}) = E(\theta_i \mid y_i) \\ &= \frac{\int (u^{y_i+1} e^{-u}/y_i!) dG(u)}{\int (u^{y_i} e^{-u}/y_i!) dG(u)} \\ &= \frac{(y_i+1) m_G(y_i+1)}{m_G(y_i)} \; , \end{aligned} \qquad (3.6)$$

where $m_G(y_i)$ is the marginal distribution of y_i. Thus, thanks to the special structure of this model, we have written the Bayes rule in terms of the data and the marginal density, which is directly estimable from the observed data.

Taking advantage of this structure, Robbins (1955) proposed the completely nonparametric estimate computed by estimating the marginal probabilities by their empirical frequencies, namely,

$$\hat{\theta}_i^B = (y_i + 1) \frac{\#(y\text{s equal to } y_i + 1)}{\#(y\text{s equal to } y_i)} \; . \qquad (3.7)$$

This formula exemplifies *borrowing information.* Since the estimate of the marginal density depends on all of the y_i, the estimate for each component θ_i is influenced by data from other components.

The foregoing is one of several models for which the analyst can make empirical Bayes inferences with no knowledge of or assumptions concerning G. Maritz and Lwin (1989) refer to such estimators as *simple EB estimators.* The added flexibility and robustness of such nonparametric procedures is very attractive, provided they can be obtained with little loss of efficiency.

The Robbins estimate was considered a breakthrough, in part because he showed that (3.7) is *asymptotically optimal* in that, as $k \rightarrow \infty$, its Bayes risk (see (B.6) in Appendix B) converges to the Bayes risk for the true Bayes rule for known G. However, in Section 3.2.3 we show that this estimator actually performs poorly, even when k is large. This poor performance can be explained in part by the estimator's failure to incorporate constraints imposed by the hierarchical structure. For example, though the Bayes estimator (3.6) is monotone in y, the Robbins estimate (3.7) need not be. Equation (3.6) also imposes certain convexity conditions not accounted for by (3.7). Several authors (van Houwelingen, 1977; Maritz and Lwin, 1989, Subsection 3.4.5) have developed modifications of the basic estimator that attempt to smooth it. One can also generalize the Robbins procedure to include models where the sampling distribution f is continuous.

Large-sample success of the Robbins approach and its refinements depends strongly on the model form and use of the posterior mean (the Bayes rule under squared error loss). A far more general and effective approach adopts the parametric EB template of first estimating G by $\widehat{G}$ and then using (3.6) with G replaced by $\widehat{G}$. That is, we use $m_{\hat{G}}(\cdot)$ in place of $\widehat{m}_G(\cdot)$. This approach accommodates general models and loss functions, imposes all necessary constraints (such as monotonicity and convexity), and provides a unified approach for discrete and continuous f. To be fully nonparametric with respect to the prior, the approach requires a fully nonparametric estimate of G. Laird (1978) proved the important result that the G which maximizes the likelihood (3.5) is a discrete distribution with at most k mass points; $\widehat{G}$ is discrete even if G is continuous. We introduce this nonparametric maximimum likelihood (NPML) estimate in Subsection 3.4.3, and consider it in more detail in Section 7.1.

3.2.3 Example: Accident data

We consider counts of accident insurance policies reporting y_i claims during a particular year (Table 3.1, taken from Simar, 1976). The data are discrete, and the Poisson likelihood with individual accident rates drawn from a prior distribution (producing data exhibiting extra-Poisson variation) is a good candidate model.

We compute the empirical Bayes, posterior mean estimates of the rate parameters θ_i using several priors estimated from the data and the "Robbins rule" that directly produces the estimates. The table shows the Robbins simple EB estimate, as well as the NPEB estimate obtained by plugging the NPML for G into the Bayes rule. As mentioned in Section 3.1, another option is to use a parametric form for G, say $G(\theta_i|\eta)$, and estimate only the parameter η. Given our Poisson likelihood, the conjugate $G(\alpha, \beta$ prior is the most natural choice Proceeding in the manner of Example 2.5, we can then estimate $\eta = (\alpha, \beta)$ from the. marginal distribution $m(\mathbf{y}|\alpha, \beta)$ plug them into the formula for the Bayes rule, and produce a parametric EB

y_i	count	$\hat{m}_G(y_i)$	$E_G(\theta_i \mid y_i)$ Robbins	Gamma	NPML
0	7840	.82867	.168	.159	.168
1	1317	.13920	.363	.417	.372
2	239	.02526	.527	.675	.610
3	42	.00444	1.333	.933	1.001
4	14	.00148	1.429	1.191	1.952
5	4	.00042	6.000	1.449	2.836
6	4	.00042	1.750	1.707	3.123
7	1	.00011	0.000	1.965	3.142

Table 3.1 *Simar (1976) Accident Data: Observed counts and empirical Bayes posterior means for each number of claims per year for k = 9461 policies issued by La Royal Belge Insurance Company. The y_i are the observed frequencies, $\hat{P}_G$ is the observed relative frequency, "Robbins" is the Robbins NPEB rule, "Gamma" is the PEB posterior mean estimate based on the Poisson/gamma model, and "NPML" is the posterior mean estimate based on the EB rule for the nonparametric prior.*

estimate. The specifics of this approach are left as Exercise 9. (The student may well wish to study Section 3.3 before attempting this exercise.)

Table 3.1 reports the empirical relative frequencies for the various observed y_i values, as well as the results of modeling the data using the Robbins rule, the PEB Poisson/gamma model, and the NPML approach. Despite the size of our dataset (k = 9461), the Robbins rule performs erratically for all but the smallest observed values of y_i. It imposes no restrictions on the estimates, and fails to exploit the fact that the marginal distribution of $\mathbf{y}$ is generated by a two-stage process with a Poisson likelihood. By first estimating the prior, either parametrically or nonparametrically, constraints imposed by the two-stage process (monotonicity and convexity of the posterior mean for the Poisson model) are automatically imposed. Furthermore, the analyst is not restricted to use of the posterior mean.

The marginal method of moments estimated mean and variance for the gamma prior are 0.2144 and 0.0160, respectively. Because the gamma distribution is conjugate, the EB estimate is a weighted average of the data y and the prior mean, with a weight of 0.7421 on the prior. For small values of y_i, all three estimates of $E(\theta_i|y_i)$ in Table 3.1 are in close agreement. The predicted accident rate for those with no accidents y_i =0) is approximately 0.16, whereas the estimate based on the MLE is 0, since $\hat{\theta}_i^{MLE} = y_i$. The Poisson/gamma and NPML methods part company for y values greater than 3. Some insight into the plateau in the NPML estimates for larger y is provided by the estimated prior $\hat{g}$, displayed in Table 3.2. Virtually all of the mass is on "safe" drivers (the first two mass points).

mass point, θ	0.089	0.580	3.176	3.669
mass, $\hat{g}(\theta)$	.7600	.2362	.0037	.0002

Table 3.2 *NPML prior estimate for the Simar Accident Data*

As y increases, the posterior loads mass on the highest mass point (3.669), but the posterior mean can never go beyond this value.

Of course, Table 3.1 provides no information on the inferential performance of these rules. Maritz and Lwin (1989, p. 86) provide simulation comparisons for the Robbins, Poisson/gamma and several "improved Bobbins" rules (they do not consider the NPML), showing that the EB approach is very effective. Tomberlin (1988) uses empirical Bayes to predict accident rates cross-classified by gender, age, and rating territory. He shows that the approach is superior to maximum likelihood in predicting the next year's rates, especially for data cells with a small number of accidents.

3.3 Parametric EB (PEB) point estimation

Parametric empirical Bayes (PEB) models use a family of prior distributions $G(\theta|\eta)$ indexed by a low-dimensional parameter η. In the context of our compound sampling framework (3.4), if η were known, this would imply that the posterior for θ_i depends on the data only through y_i, namely,

$$p(\theta_i|y_i,\eta) = \frac{f_i(y_i|\theta_i)g(\theta_i|\eta)}{m_i(y_i|\eta)} . \tag{3.8}$$

But since η is unknown, we use the marginal distribution of all the data $m(\mathbf{y}|\eta)$ to compute an estimate $\hat{\eta}$, usually obtained as an MLE or method of moments estimate. Plugging this value into (3.8), we get the *estimated* posterior $p(\theta_i|y_i,\hat{\eta})$. This estimated posterior drives all Bayesian inferences; for example, a point estimate $\hat{\theta}_i$ is the estimated posterior mean (or mode, or median). Note that $\hat{\theta}_i$ depends on *all* the data through $\hat{\eta} = \hat{\eta}(\mathbf{y})$ Morris (1983a) and Casella (1985) provide excellent introductions to PEB analysis.

3.3.1 *Gaussian/Gaussian models*

We now consider the two-stage Gaussian/Gaussian model,

$$\begin{aligned} Y_i \mid \theta_i &\stackrel{iid}{\sim} N(\theta_i,\sigma^2),\ \ i=1,\ldots,k, \text{ and} \\ \theta_i \mid \mu &\stackrel{iid}{\sim} N(\mu,\tau^2),\ \ i=1,\ldots,k. \end{aligned} \tag{3.9}$$

First, we assume both τ^2 and σ^2 are known, so that $\eta = \mu$. Calculations similar to those for the posterior density in Example 2.1 show that the Y_i are marginally independent and identically distributed as $N(\mu,\ \sigma^2+\tau^2)$

random variables. Hence (3.5) implies that

$$m(\mathbf{y}|\mu) = \frac{1}{[2\pi(\sigma^2+\tau^2)]^{n/2}} \exp\left[-\frac{1}{2(\sigma^2+\tau^2)}\sum_{i=1}^{k}(y_i-\mu)^2\right] ,$$

so that $\hat{\mu} = \bar{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$ is the marginal MLE of μ. Thus the estimated posterior distribution is

$$p(\theta_i|y_i,\hat{\mu}) = N(B\hat{\mu}+(1-B)y_i \ , \ (1-B)\sigma^2) \ ,$$

where $B = \sigma^2/(\sigma^2+\tau^2)$. This produces the PEB point estimate of θ_i:

$$\hat{\theta}_i^{\mu} = B\bar{y}+(1-B)y_i \quad (3.10)$$

$$= \bar{y}+(1-B)(y_i-\bar{y}) \ . \quad (3.11)$$

Notice that this formula is identical to that derived in Example 2.1, except that a known prior mean (μ) is replaced by the sample mean using all of the data. Inference on a single component (a single θ_i depends on data from all the components (borrowing information). Formula (3.10) shows the estimate to be a weighted average of the estimated prior mean $(\bar{y})$ and the standard estimate (y_i), while (3.11) shows that the standard estimate is shrunk back toward the common mean a fraction of the distance between the two. The MLE is equivalent to setting $B = 0$, while complete pooling results from setting $B = 1$.

As in the Robbins example, the PEB estimates borrow information from all the coordinates in making inference about a single coordinate. In a slightly more general setting below, we shall show that these PEB estimates outperform the usual MLE when the two-stage model holds. This result is not surprising, for we are taking advantage of the connection between the k coordinate-specific θs provided by *G*. What is surprising is that estimates such as those above can outperform the MLE even when *G* is misspecified, or even when the θs are not sampled from a prior.

We now generalize the above by retaining (3.9), but assuming that *r* is also unknown. The marginal likelihood must now provide information on $\eta = (\mu,\tau)$. We are immediately confronted with a decision on what estimate to use for τ (or τ^2 or *B*). If we use the marginal MLE (MMLE) for τ^2, then the invariance property of MLEs implies that the MMLE for *B* results from plugging $\hat{\tau}^2$ into its formula. For model (3.9), it is easy to show that the MMLE for μ is again $\bar{y}$, while the MMLE for τ^2 is

$$\hat{\tau}^2 = (s^2-\sigma^2)^+ = \max\{0, s^2-\sigma^2\} \ , \quad (3.12)$$

where $s^2 = \frac{1}{k}\sum_{i=1}^{k}(y_i-\bar{y})^2$. This "positive part" estimator (3.12) reports the variation in the data over and above that expected if all the θs are equal (i.e., σ^2). Transforming (3.12), we find that the MMLE for *B is*

$$\widehat{B} = \frac{\sigma^2}{\sigma^2+\hat{\tau}^2} = \frac{\sigma^2}{\sigma^2+(s^2-\sigma^2)^+} \ . \quad (3.13)$$

Clearly $0 \leq \widehat{B} \leq 1$; this would not have been the case had we used the unbiased estimate of τ^2 which divides $\sum_{i=1}^{k}(y_i - \bar{y})^2$ by k - 1 instead of k and does not take the positive part. Since we know the target quantity B must lie in [0, 1], this provides another illustration of the problems with unbiased estimates.

Substituting (3.13) for B in (3.11) produces the EB estimates

$$\hat{\theta}_i^{\mu\tau} \quad = \quad \bar{y} + (1 - \widehat{B})(y_i - \bar{y}) \, . \tag{3.14}$$

As in the case where τ is known, the target for shrinkage is $\bar{y}$, but now the amount of shrinkage is controlled by the estimated heterogeneity in the data. The more the heterogeneity exceeds σ^2, the smaller $\widehat{B}$ becomes, and the less the shrinkage. On the other hand, if the estimated heterogeneity is less than σ^2, $\widehat{B} = 1$ and all estimates equal $\bar{y}$ (i.e., the data are pooled). As heterogeneity increases, the weight on the estimated coordinate-specific estimates increases to 1.

As discussed by Laird and Louis (1987), representation (3.11) shows a direct analogy with regression to the mean. In the model, Y_i and θ_i have correlation (the infra-class correlation) equal to (1 - B), and an estimate for θ_i based on Y_i regresses back toward the common mean. In the usual regression model, we estimate the means and correlations from a data set where both the Ys and θs (the response variables) are observed. In Bayes and empirical Bayes analysis, a distribution on the θs provides sufficient information for inference even when the θs are not observed.

Example 3.1 We apply our Gaussian/Gaussian EB model (3.9) to the data in Table 3.3. Originally analyzed in Louis (1984), these data come from the Hypertension Detection and Follow-up Program (HDFP). This study randomized between two intervention strategies: referred care (RC), in which participants were told they had hypertension (elevated blood pressure) and sent back to their physician, and stepped care (SC), in which participants were entered into an intensive and hierarchical series of lifestyle and treatment interventions. The data values Y are MLE estimates of 1000 times the log-odds ratios in comparing five-year death rates between the SC and RC regimens in twelve strata defined by initial diastolic blood pressure (I = 90-104, II = 105-114, 111 = 115+), race (B/W), and gender (F/M).

In a stratum with five-year death rates r_{sc} and r_{rc}, we have

$$Y = 1000 \times \log\left(\frac{r_{sc} \cdot (1 - r_{rc})}{(1 - r_{sc}) \cdot r_{rc}}\right) \; .$$

These estimated log-relative risks have estimated sampling standard deviations which we denote by $\hat{\sigma}$. These vary from stratum to stratum, and control the amount of shrinkage in a generalization of the empirical Bayes rule (3.14) wherein $\hat{B}$ is replaced by $\hat{B}_i = \hat{\sigma}_i^2/(\hat{\sigma}_i^2 + \hat{\tau}^2)$. In this unequal sampling variance situation, MMLEs for μ and τ^2 require an iterative algorithm (see Subsection 3.4).

Group		Y	$\hat{\theta}^{\mu\tau}$	D	$\hat{\sigma}$	PSD	$\hat{\theta}^{con}$	A
I	BM	−129	−157	54	170	125	−149	81
	BF	−304	−240	44	206	137	−285	74
	WM	−242	−220	59	153	117	−240	84
	WF	−355	−253	39	231	144	−316	69
II	BM	−274	−213	29	290	155	−255	59
	BF	−529	−266	23	337	161	−383	51
	WM	−41	−156	22	349	162	−136	50
	WF	809	−61	13	479	171	127	34
III	BM	−558	−273	23	337	161	−398	51
	BF	−235	−197	18	389	166	−231	44
	WM	336	−122	13	483	171	−38	34
	WF	1251	−103	6	730	178	43	18

Table 3.3 *Analysis of 1000 x log(odds ratios) for five-year survival in the HDFP study:* $\hat{\mu}$ = *-188,* $\hat{\tau}^2 = (183)^2$ $D = 100(1 - B)$*, and PSD* = $\hat{\sigma}D^{\frac{1}{2}}/10$*, the posterior standard deviation.* $\hat{\theta}^{\mu\tau}$ = *-188 +* $\frac{D}{100}(Y + 188)$*, the usual PEB point estimate, while* $\hat{\theta}^{con}$ = *-229* $\frac{A}{100}(Y + 229$*, the Louis (1984) constrained EB estimate.*

Table 3.3 displays both the EB posterior means, $\hat{\theta}^{\mu\tau}$ and a set of *constrained* EB estimates, $\hat{\theta}^{con}$. Like the usual EB estimate, the constrained estimate is a weighted average of the estimated prior mean and the data, but with an adjusted prior mean estimate (here, -229 instead of -188) and a different weight on the data ($A/100$). Constrained estimates are designed to minimize SEL subject to the constraint that their sample mean and variance are good estimates of their underlying empirical distribution function of the θs. We defer the derivation and specifics of constrained EB estimation and histogram estimates more generally to Section 7.1.

Notice from Table 3.3 and Figure 3.1 that the subgroup estimates with high variance for the MLE tend to produce the most extreme observed estimates and are shrunken most radically toward the estimated population mean (i.e., the value of 1- $\widehat{B}$ is relatively small). The PSD column reports "naive" posterior standard errors that have not been adjusted for having estimated parameters of the prior. Notice that these are far more homogeneous than the $\hat{\sigma}$ values; highly variable estimates have been stabilized. However, as we shall show in Section 3.5, these values need to be increased to account for having estimated the prior parameters. This increase will be greater for subgroups with a smaller value of 1 - B and for those further from $\hat{\mu}$.

Typical relations between observed data and empirical Bayes estimates are shown in Figure 3.1, which plots the raw Y values on the center axis,

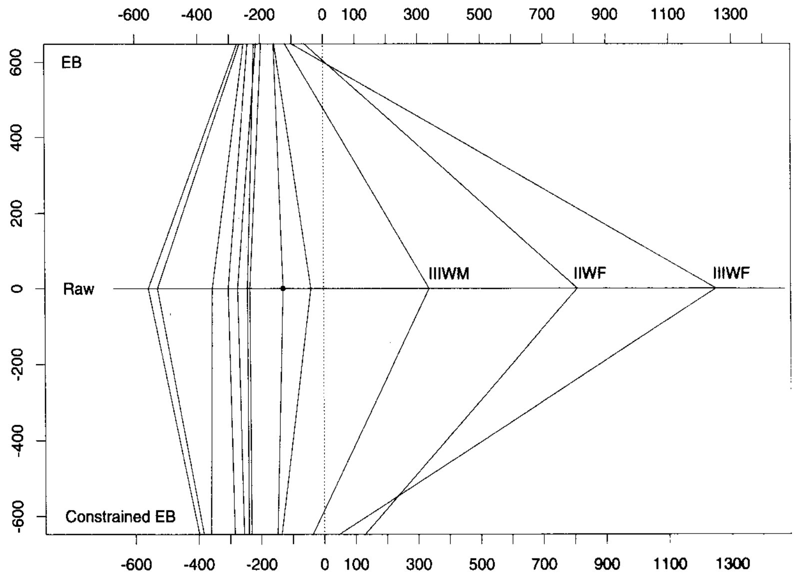


Figure 3.1 *HDFP raw data and EB estimates.*

connected by solid lines to the corresponding EB estimates $\hat{\theta}^{\mu\tau}$ on the upper axis. The most extreme estimates are shrunk the most severely, and rankings of the estimates can change (i.e., the lines can cross) due to the unequal sampling variances of the Ys. Efron and Morris (1977) give an example of this phenomenon related to geographic distribution of disease. The constrained EB estimates $\hat{\theta}^{con}$ plotted on the figure's lower axis, shrink observed estimates less than the posterior means, and produce ranks different from either the raw data or the standard EB estimates. ■

The Gaussian/Gaussian model exhibits the more general phenomenon that EB estimates provide a compromise between pooling all data (and thus decreasing variance while increasing the bias of each coordinate-specific estimate) and using only coordinate-specific data for the coordinate-specific estimates (and increasing variance while decreasing bias). Complete pooling occurs when $\hat{B} = 1$, i.e., when the chi-square statistic for testing equality of the k means is less than or equal to $k/(k - 1)$ or large k and $\tau = 0$, the probability of complete pooling (equivalent to accepting the hypothesis that the coordinate-specific means are equal) is approximately .50. Thus, the test has type I error of .50, but this apparent liberality is compensated for by the smooth departure from complete pooling.

Proper selection of hyperparameter estimates is one inherent difficulty with the PEB approach. We have already proposed that using the positive part estimator for B in (3.13) is more sensible than using an unbiased estimate, but many other options for $\widehat{B}$ are possible, and it may be difficult to guess which will be best in a given situation. By contrast, the fully Bayesian model that adds another level to the distributional hierarchy (i.e. including a hyperprior $h(\mu, \tau^2)$) replaces estimation by integration, thus avoiding the problem of selecting an estimation criterion. This solution also deals effectively with the non-identically distributed case (for example, when the sampling variances are different), and automatically injects uncertainty induced by estimating the hyperparameters into the posterior distribution. However, this does require selection of an effective hyperprior, an issue we shall revisit in Subsection 3.5.2 and the subsequent Example 3.3.

3.3.2 Beta/binomial model

As a second example, consider the EB approach to estimating the success probability in a binomial distribution. With X the number of events in n trials and θ the event probability, the sampling distribution is

$$P(X = x \mid \theta) = f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} .$$

For the empirical Bayesian analysis we follow Subsection 2.3.4 and use the conjugate $Beta(a, b)$ prior distribution. For convenience, we reparametrize from (a, b) to $(\mu\ M)$ where $\mu = a/(a+b)$, the prior mean, and $M = a + b$, a measure of prior precision (i.e., increasing M implies decreasing prior variance). Specifically, $V(\theta \mid \mu,\ M) = \mu(1-\mu)/(M+1)$.

The marginal distribution of X is beta-binomial with density

$$P(X = x \mid \mu,\ M) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))}\binom{n}{x}\frac{\Gamma(x+M\mu)\Gamma(n-x+M(1-\mu))}{\Gamma(n+M)} . \tag{3.15}$$

Notice that if $\mu = .5$ and $M = 2$, θ is uniformly distributed *a priori*, and the marginal distribution is $P(X = x \mid \mu,\ M) = 1/(n+1)$ a discrete uniform distribution on the integers 0, 1, . . ., n.

In the EB context we have a two-stage, compound sampling model where $\theta_i \overset{iid}{\sim} Beta(\mu, M)$, and $X_i|\theta_i \sim Bin(n_i,\ \theta_i)$. The MMLE for μ, M), or equivalently (a, b), *is* found by subscripting n and x by i and maximizing the product of (3.15) over i. This maximization requires evaluation of the gamma and digamma functions. A method of moments estimate, while not as efficient as the MMLE, is revealing. Using the iterated expectation and variance formulae, and denoting the prior mean and variance of θ by μ and τ^2 we have

$$E\left(\frac{X_i}{n_i}\right) = E\left[E\left(\frac{X_i}{n_i}\Big|\theta_i\right)\right] = E\left[\theta_i\right] = \mu$$

and

$$\begin{aligned}
V\left(\frac{X_i}{n_i}\right) &= E\left[V\left(\frac{X_i}{n_i}\middle|\theta_i\right)\right] + V\left[E\left(\frac{X_i}{n_i}\middle|\theta_i\right)\right] \\
&= E\left[\frac{\theta_i(1-\theta_i)}{n_i}\right] + V\left[\theta_i\right] \\
&= \frac{\mu(1-\mu)}{n_i} + \tau^2\frac{n_i - 1}{n_i} \\
&= \frac{\mu(1-\mu)}{n_i}\left[1 + \frac{n_i - 1}{M+1}\right] .
\end{aligned}$$

The pooled mean, X_+/n_+ is a natural method of moments estimate for μ, but with unequal n_i, there are a variety of method of moments estimates for μ and M (see Louis and DerSimonian, 1982). If all the n_i are equal, matching $s^2 = \frac{1}{k}\sum_i(\frac{X_i}{n} - \hat{\mu})^2$ and the marginal variance produces

$$\widehat{M} = \frac{\hat{\mu}(1-\hat{\mu}) - s^2}{s^2 - \frac{\hat{\mu}(1-\hat{\mu})}{n}}. \tag{3.16}$$

Notice that $\widehat{M}$ decreases to 0 as s^2 increases to its maximum value of $\hat{\mu}(1-\hat{\mu})$, and increases to infinity as s^2 decreases to its binomial distribution value, $\hat{\mu}(1-\hat{\mu})/n$.

The estimates $\hat{\mu}$ and $\widehat{M}$ can now be used in a plug-in EB approach. The estimated conditional distribution of θ_i given X_i is $Beta(X_i + \widehat{M}\hat{\mu},\ n_i - X_i + \widehat{M}(1-\hat{\mu}))$. The estimated posterior mean is then

$$\begin{aligned}
\theta_i^{\hat{\mu},\widehat{M}} &= \hat{\mu}_i = \frac{\widehat{M}\hat{\mu} + X_i}{\widehat{M} + n_i} \\
&= \frac{\widehat{M}}{\widehat{M} + n_i}\hat{\mu} + \frac{n}{\widehat{M} + n_i}\frac{X_i}{n_i} ,
\end{aligned}$$

which is the posterior mean with "hatted" values for the prior parameters. Hence $\theta_i^{\hat{\mu},\widehat{M}}$ is a weighted average of the prior mean and the MLE, X_i/n_i, with weight depending on the relative size of $\widehat{M}$ and n_i. Increasing $\widehat{M}$ (decreasing s^2) puts more weight on the prior mean.

The estimated posterior variance is $[\hat{\mu}_i(1-\hat{\mu}_i)]/(\widehat{M} + n_i)$. As we warned for the Gaussian/Gaussian model, confidence intervals based on this variance or on the full estimated posterior will be liberal since they do not account for uncertainty in estimating the prior parameters.

For an example of applying the EB approach to inference for binomial distributions, see Tarone (1982) and Tamura and Young (1986). A fully Bayesian alternative is provided by Dempster et al. (1983).

3.3.3 EB Performance of the PEB

To see how well the empirical Bayes rule works when model (3.9) holds, we can compute the preposterior (Bayes) risk (B.6) for a single component or for the average over all components. Since the Y_i are i.i.d., the two computations will be equal and we take advantage of this equality.

Calculating the preposterior risk requires integration over both the data and parameters, and we may integrate in either order. As shown in Appendix B, integrating only over the distribution of the parameters conditional on the data gives the posterior risk, while integrating only over the sampling distribution of the data conditional on the parameters gives the frequentist risk. Each is revealing. First, we consider the posterior risk for the i^{th} component of the EB estimator (3.14). Under squared error loss, the usual posterior "variance plus squared bias" formula gives

$$(1-B)\sigma^2 + [B\mu + (1-B)y_i - \widehat{B}\bar{y} - (1-\widehat{B})y_i]^2 .$$

Regrouping terms gives

$$(1-B)\sigma^2+B^2(\bar{y}-\mu)^2+(\widehat{B}-B)^2(y_i-\bar{y})^2-2B(\widehat{B}-B)(y_i-\bar{y})(\bar{y}-\mu). \quad (3.17)$$

The first term is the usual Bayes posterior risk. The second term accounts for estimating μ, the third for estimating B, and the fourth for correlation in these estimates. Notice that the last two depend on how far y_i is from $\bar{y}$, and that the second and third are very similar to those used to compute confidence intervals for regression lines (see, e.g., Snedecor and Cochran, 1980, Section 9.9).

To get the preposterior risk we take the expectation of (3.17) with respect to the marginal distribution of $\mathbf{Y}$. The computation is far easier if we first average over all coordinates, to obtain

$$\frac{1}{k}\sum_{i=1}^{k} E((\theta_i - \hat{\theta}_i^{\mu\tau})^2 \mid \mathbf{y}) = (1-B)\sigma^2 + B^2(\bar{y}-\mu)^2 + (\widehat{B}-B)^2 SS_y, \quad (3.18)$$

where $SS_y = \sum_{i=1}^{k}(y_i - \bar{y})^2$ since the last term in (3.17) sums to 0.

The marginal distribution of each Y_i is Gaussian with mean μ and variance $\sigma^2 + \tau^2 = \sigma^2/B$. For ease of calculation we take the estimate of B to be of the form

$$\widehat{B} = \frac{c\sigma^2}{SS_y} \quad (3.19)$$

(i.e., we do not use a "positive part," as in (3.13)). We then expand the third term in (3.18) and take expectations using properties of the gamma distribution to obtain the preposterior risk,

$$r(G, \hat{\theta}^{\mu\tau}) = (1-B)\sigma^2 + \frac{B\sigma^2}{k} + \frac{B\sigma^2}{k}\left(\frac{c^2}{k-3} - 2c + k - 1\right) \quad (3.20)$$

provided that $k \geq 4$. Again, the first term is the usual Bayes risk, the second accounts for estimating μ, and the third accounts for estimating B. The unbiased estimate of B uses $c = k - 3$, giving $2B\sigma^2/k$ for the last term. This is also its minimum value for this restricted class of estimators, but overall performance can be improved by improving the estimate of B (for example, by truncating it at 1).

Formula (3.20) shows the Bayes/EB advantage, when the compound sampling model holds. The usual MLE $(\hat{\theta} = y_i)$ has preposterior risk σ^2, so for $c = k - 3$, the savings in risk is

$$B\sigma^2 \left(1 - \frac{3}{k}\right) ,$$

which can be substantial. Recall that this savings applies to each coordinate θ_i, though we have averaged over coordinates in our calculations.

3.3.4 Stein estimation

The above results show that there is great potential to improve on the performance of the standard maximum likelihood estimate when the two-stage structure holds and the assumed form of the prior distribution is correct. The prior provides a connection among the coordinate-specific parameters.

Now consider the other extreme where there is no prior distribution: we assume only that $Y_i|\theta_i \stackrel{ind}{\sim} N(\theta_i, \sigma^2)$, $i = 1, \ldots, k$, where σ^2 is assumed known. We investigate the purely frequentist properties of the EB procedure for the fixed vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. For there to be any hope of improvement over the maximum likelihood estimate $\hat{\boldsymbol{\theta}}^{MLE}(\mathbf{Y}) = \mathbf{Y}$, we must provide a connection among the coordinates, and, without a prior, we do this on the "other end" by assessing *summed squared error doss* (SSEL),

$$SSEL(\mathbf{Y}, \boldsymbol{\theta}) = \frac{1}{k} \sum_{i=1}^{k} (\hat{\theta}_i(\mathbf{Y}) - \theta_i)^2 . \tag{3.21}$$

Hence, while the parameters are unconnected in the model, our inferences regarding them *are* connected through the loss function. Note that for convenience, we are actually computing *averaged* squared error loss, but will still refer to it as SSEL.

This loss structure is "societal" in that it combines performance over all coordinates and does not pay specific attention to a single coordinate as would, say, the maximum of the coordinate-specific losses. The results that follow are dependent on a loss function that averages over coordinates; the data analyst must decide if such a loss function is relevant.

Consider the frequentist risk (equation (B.4) in Appendix B) for SSEL,

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{Y})) = E(SSEL(\mathbf{Y}, \boldsymbol{\theta}) \mid \boldsymbol{\theta}) , \tag{3.22}$$

and our Gaussian sampling distribution. It is easy to see that

$$\begin{aligned} R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{MLE}(\mathbf{Y})\right) &= \frac{1}{k}\sum_{i=1}^{k} E((Y_i - \theta_i)^2|\boldsymbol{\theta}) \\ &= \frac{1}{k}\sum_{i=1}^{k} \sigma^2 \;=\; \sigma^2 \,, \end{aligned}$$

which, of course, is constant regardless of the true value of $\boldsymbol{\theta}$. Stein (1955) proved the remarkable result that for $k \geq 3$, $\hat{\boldsymbol{\theta}}^{MLE}(\mathbf{Y})$ *is inadmissible* as an estimator of $\boldsymbol{\theta}$. That is, there exists another estimator with frequentist risk no larger than σ^2 for every possible $\boldsymbol{\theta}$ value. One such dominating estimator was derived by James and Stein (1961) as

$$\hat{\theta}_i^{JS}(\mathbf{Y}) = \left[1 - \frac{(k-2)\sigma^2}{||\mathbf{Y}||^2}\right] Y_i \,, \tag{3.23}$$

where $||\mathbf{Y}||^2 \equiv \sum Y_i^2$. These authors showed that

$$\begin{aligned} R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{JS}(\mathbf{Y})\right) &= \frac{1}{k}\sum_{i=1}^{k} E((\hat{\theta}_i^{JS} - \theta_i)^2|\boldsymbol{\theta}) \\ &= \sigma^2 \left[1 - \frac{k-2}{k} E_{\mathbf{Y}|\boldsymbol{\theta}}\left(\frac{(k-2)\sigma^2}{||\mathbf{Y}||^2}\right)\right] \,, \end{aligned} \tag{3.24}$$

which is less than σ^2 provided $k > 3$. The subscript on the expectation in (3.24) indicates that it is with respect to the full sampling distribution $f(\mathbf{Y}|\boldsymbol{\theta})$, but it actually only requires knowledge of the simpler, univariate distribution of $||\mathbf{Y}||$ given $||\boldsymbol{\theta}||$.

The risk improvement in (3.24) is often referred to as "the Stein effect." While it is indeed surprising given the mutual independence of the k components of $\mathbf{Y}$ it does depend strongly on a loss function that combines over these components. The reader may also wonder about its connection to empirical Bayes analysis. This connection was elucidated in a series of papers by Efron and Morris (1971, 1972a,b, 1973a,b, 1975, 1977). Among many other things, these authors showed that $\hat{\boldsymbol{\theta}}^{JS}$ is exactly the PEB point estimator in the Gaussian/Gaussian model (3.9) that assumes μ to be known and equal to 0, and estimates B by

$$\widehat{B} = \frac{k-2}{||\mathbf{Y}||^2} \,, \tag{3.25}$$

an unbiased estimator of B (the proof of this fact is left as an exercise). As suggested by the previous subsection, EB performance of the James-Stein estimator can be improved by using a positive part estimator of B, leading to

$$\hat{\theta}_i^{JS+}(\mathbf{Y}) = \left[1 - \min\left(\frac{(k-2)\sigma^2}{||\mathbf{Y}||^2}\,,\,1\right)\right] Y_i\ ,$$

so that estimates of θ_i are never shrunk *past* the prior mean, 0. (Notice that this improvement implies that the original James-Stein estimator is itself inadmissible!) Another way to generalize $\hat{\theta}_i^{JS}(\mathbf{Y})$ is to shrink toward prior means other than $\mu = 0$, or toward the *sample* mean in the absence of a reliable prior estimate for μ. This latter estimator is given by

$$\hat{\theta}_i^{JS'}(\mathbf{Y}) = \bar{Y} + \left[1 - \frac{(k-3)\sigma^2}{||\mathbf{Y} - \bar{Y}||^2}\right](Y_i - \bar{Y})\ .$$

Note that this estimator is a particular form of our Gaussian/Gaussian EB point estimator (3.14). For SSEL, one can show that

$$R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{JS'}(\mathbf{Y})\right) = \sigma^2\left[1 - \frac{k-3}{k} E_{\mathbf{Y}|\boldsymbol{\theta}}\left(\frac{k-3}{||\mathbf{Y} - \bar{Y}||^2}\right)\right]\ ,$$

where the expectation depends only on $||\boldsymbol{\theta} - \bar{\theta}||^2$. Hence $\hat{\boldsymbol{\theta}}^{JS'}(\mathbf{Y})$ dominates the MLE $\mathbf{Y}$ provided $k \geq 4$, a slightly larger dataset being required to "pay" for our estimation of μ.

The basic result (3.24) can be proven directly (see, e.g., Cox and Hinkley, p. 446-8), or indirectly by showing that the expectation of (3.24) when the θ_i are i.i.d. from a $N(0, \tau^2)$ prior matches the preposterior risk of $\hat{\boldsymbol{\theta}}^{JS}$ under this prior for all r. But these expectations depend only on $||\boldsymbol{\theta}||^2$, which is distributed as a multiple of a central chi-square distribution. Completeness of the chi-square distribution dictates that, almost surely, there can be only one solution to this relation, and the Stein result follows. We leave the details as an exercise.

Figure 3.2 plots the frequentist risk of the James-Stein estimator when $\sigma^2 = 1$ and $k = 5$. As $||\theta||^2$ increases, the risk improvement relative to the MLE decreases and eventually disappears, but performance is never worse than the MLE $\mathbf{Y}$. Since $\mathbf{Y}$ is also the minimax rule, the James-Stein estimator is *subminimax* for $k \geq 3$ (see Subsection B.1.4 in Appendix B).

Coordinate-specific vs. coordinate average loss

The PEB results in Subsection 3.3.3 show how effective the empirical Bayes approach can be in estimating individual parameters when the assumed compound structure holds. The Stein results in Subsection 3.3.4 show that the EB estimators perform well for all $\boldsymbol{\theta}$ vectors, when evaluated under SSEL. However, the coordinate-specific risk of these estimators can be high if a relatively small number of the θs are outliers (an unlikely event if the

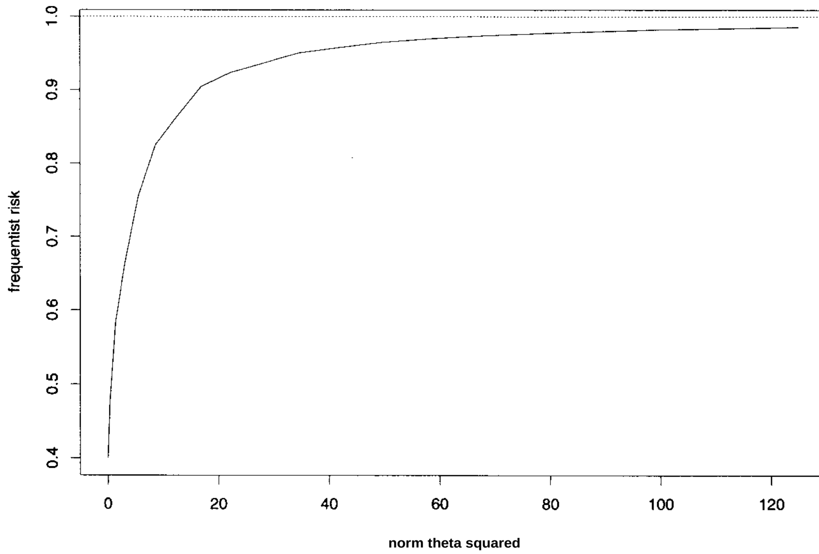


Figure 3.2 *Frequentist risk for Stein's rule with $k = 5$*

θ_i are sampled from a Gaussian prior). In fact, one can show that

$$\sup_{\boldsymbol{\theta}} \max_{1 \leq i \leq k} R(\theta_i, \hat{\theta}_i^{JS}) = \frac{k\sigma^2}{4}, \tag{3.26}$$

where $R(\theta_i, \hat{\theta}_i^{JS})$ is defined as a single component of the sum in (3.24). The supremum is attained by setting $\theta_1 = \theta_2 = \cdots = \theta_{k-1} = 0$ and letting $\theta_k \to \infty$ at rate $\sqrt{k}$ (see the exercises). In this limiting case, the *average* expected squared error loss is $\frac{\sigma^2}{4}$, a substantial improvement over the MLE. But, for the MLE, the supremum is σ^2 and Stein's rule can produce a substantial coordinate-specific penalty in this extreme situation. This poor performance for outlying parameters has motivated *limited translation rules* (Efron and Morris, 1972a), which are compromises between MLEs and Stein rules that limit componentwise risk. Selection of priors and likelihoods which are more robust against outliers is another possible remedy, and one we shall return to in Chapter 6.

In most applications, even if the assumed prior does not hold, the coordinate parameters are not so completely unrelated as in the situation producing (3.26), and so, when carefully applied, the empirical Bayes approach is effective even for coordinate-specific evaluations. Many empirical case studies support this point; for a sampler of applications, see Rubin (1980) on law school admissions, Tomberlin (1988) on insurance rating, Efron and Morris (1973b, 1977) on batting averages and toxoplasmosis, Devine et al. (1994, 1996) on disease mapping and accident mapping, van Houwelingen and Thorogood (1995) on kidney graft survival modeling, and Zaslavsky (1993) on the census undercount.

In other applications, SSEL may be the most reasonable performance benchmark. For example, Hui and Berger (1983) use the approach to stabilize residual variance estimates based on a small number of degrees of freedom. The estimated variances are then used to produce weights in a subsequent analysis. Overall performance is important here, with little interest in individual variance estimates.

3.4 Computation via the EM algorithm

For the two-stage model, EB analysis requires maximization of the marginal likelihood given in (3.2). For many models, standard iterative maximum likelihood methods can be used directly to produce the MMLE and the observed information matrix. In more complicated settings, the *EM algorithm* (where EM stands for "expectation-maximization") offers an alternative approach, as we now describe.

Versions of the EM algorithm have been used for decades. Dempster, Laird, and Rubin (1977) consolidated the theory and provided instructive examples. The EM algorithm is attractive when the function to be maximized can be represented as a missing data likelihood, as is the case for (3.2). It converts such a maximum likelihood situation into one with a "pseudo-complete" log-likelihood or score function, produces the MLE for this case, and continues recursively until convergence. The approach is most effective when finding the complete data MLE is relatively straightforward. Several authors have shown how to compute the observed information matrix within the EM structure, or provided various extensions intended to accelerate its convergence or ease its implementation.

Since the marginal likelihood (3.2) can be represented as a missing data likelihood, the EM algorithm is effective. It also provides a conceptual introduction to the more advanced Monte Carlo methods for producing the full posterior distribution presented in Section 5.4. Here we sketch the approach and provide basic examples.

3.4.1 EM for PEB

Consider a model where, if we observed $\boldsymbol{\theta}$, the MLE for η would be relatively straightforward. For example, in the compound sampling model (3.4), given $\boldsymbol{\theta}$ the MLE for η can be computed using only the prior g. Suppressing the dependency on $\mathbf{y}$ on the left-hand side of equations, let

$$S(\boldsymbol{\theta}|\eta) \quad = \quad \frac{\partial}{\partial\eta}\log(g(\boldsymbol{\theta}|\eta)) \tag{3.27}$$

be the score function. For the "E-step," let $\eta^{(j)}$ denote the current estimate of the hyperparameter at iteration j, and compute

$$\bar{S}(\eta|\eta^{(j)}) \quad = \quad E(S(\boldsymbol{\theta}|\eta) \,|\, \mathbf{y}, \, \eta^{(j)}) \, .$$

This step involves standard Bayesian calculations to arrive at the conditional expectation of the sufficient statistics for η based on $\boldsymbol{\theta}$.

The "M-step" then uses $\bar{S}$ in (3.28) to compute a new estimate of the hyperparameter, and the recursion proceeds until convergence. That is,

$$\eta^{(j+1)} = arg \max_{\eta} \bar{S}(\eta|\eta^{(j)}) \; , \quad j = 1, 2, \ldots \tag{3.29}$$

If both the E and M steps are relatively straightforward (the most attractive case being when the conditional distributions in (3.28) and (3.29) are exponential families), the EM approach works very well.

Dempster et al. (1977), Meilijson (1989), Tanner (1993, p.43), and many other researchers have shown that, at every iteration, the marginal likelihood either increases or stays constant, i.e.,

$$m(\mathbf{y} \mid \eta^{(j+1)}) \geq m(\mathbf{y} \mid \eta^{(j)}) \; \text{ for all } j \; ,$$

and thus the algorithm converges monotonically. This convergence could be to a local (rather than global) maximum, however, and so, as with many optimization algorithms, multiple starting points are recommended.

Example 3.2 Consider the Gaussian/Gaussian model (3.9) with unequal sampling variances, and let $\eta = (\mu, T)$, where for notational convenience we use T for τ^2. For model component i, -2 times the loglikelihood for η as a function of θ_i is log(T) + $(\theta_i - \mu)^2)/T$, producing the score vector

$$S(\theta_i|\eta) = - \begin{pmatrix} -2\frac{(\theta_i - \mu)}{T} \\ \frac{1}{T} - \frac{(\theta_i - \mu)^2}{T^2} \end{pmatrix} . \tag{3.30}$$

Expanding the second component gives

$$\frac{1}{T} - \frac{\theta_i^2 - 2\theta_i\mu + \mu^2}{T^2} \; .$$

The MLE for η depends on the sufficient statistics $\sum_i \theta_i$ and $\sum_i \theta_i^2$. The EM proceeds by computing $\bar{S}$ as in (3.28) using the conditional posterior distribution $p(\theta_i|y_i, \mu^{(j)}, T^{(j)})$, substituting the posterior mean for θ_i and the posterior variance plus the square of the mean for θ_i^2, solving the "psuedo-complete"-data MLE equations, and continuing the recursion. Note that the algorithm requires expected sufficient statistics, so θ_i^2 must be replaced by its expected value, not by the square of the expectation of θ_i (as though the imputed θs were observed). ■

For this model and generalizations, the EM partitions the likelihood maximization problem into components that are relatively easy to handle. For example, unequal sampling variances produce no additional difficulty. The

EM is most attractive when both the E and the M steps are relatively more straightforward than dealing directly with the marginal likelihood. However, even when the E and M steps are non-trivial, they may be handled numerically, resulting in potentially simplified or stabilized computations.

3.4.2 *Computing the observed information*

Though the EM algorithm can be very effective for finding the MMLE, it never operates directly on the marginal likelihood or score function. It therefore does not provide a direct estimate of the observed information needed to produce the asymptotic variance of the MMLE. Several approaches to estimating observed information have been proposed; all are based on the standard decomposition of a variance into an expected conditional variance plus the variance of a conditional expectation.

Using the notation for hierarchical models and suppressing dependence on η, the decomposition of the Fisher information for η is

$$I_{\boldsymbol{\theta}} = E(I_{\boldsymbol{\theta}|\mathbf{Y}} + I_{\mathbf{Y}}) \, , \tag{3.31}$$

where I is the information that would result from direct observation of $\boldsymbol{\theta}$. The first term on the right is the amount of information on $\boldsymbol{\theta}$ provided by the conditional distribution of the complete data given y, while the second term is the information on η in the marginal likelihood (the required information for the MMLE).

Louis (1982) estimates observed information versions of the left-hand side and the first right-hand term in (3.31), obtaining the second term by subtraction. Meng and Rubin (1991) generalize the approach, producing the *Supplemented EM* (SEM) algorithm. It requires only the EM computer code and a subroutine to compute the left-hand side. Meilijson (1989) shows how to take advantage of the special case where **y** is comprised of independent components to produce an especially straightforward computation. With *i* indexing components, this approach uses (3.28) for each component, producing $\bar{S}_i$, and then

$$\hat{I}_{\mathbf{Y}} = \sum_i \bar{S}_i \bar{S}_i^T \, , \tag{3.32}$$

requiring no additional computational burden. This gradient approach can be applied to (3.30) in the Gaussian example. With unequal sampling variances (σ_i^2), validity of the approach requires either that they are sampled from a distribution (a "supermodel"), producing an i.i.d. marginal model, or weaker conditions ensuring large sample convergence. Otherwise, the more complicated approaches of Louis (1982) or Meng and Rubin (1991) must be used.

3.4.3 EM for NPEB

The EM algorithm is also ideally suited for computing the nonparametric maximum likelihood (NPML) estimate, introduced in Subsection 3.2 above. To initiate the EM algorithm, recall that the prior $G(\boldsymbol{\theta})$ which maximizes the likelihood (3.5) in the nonparametric setting must be a discrete distribution with at most k mass points. Thus, assume the prior has mass points $\boldsymbol{\mu}^{(0)} = (\mu_1^{(0)}, \ldots, \mu_J^{(0)})$ and probabilities $\boldsymbol{\pi}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_J^{(0)})$, where $J = k$. Then, for the $(\nu+1)^{st}$ iteration, let

$$\begin{aligned} w_{ij}^{(\nu+1)} &= P(\theta_i = \mu_j^{(\nu)} \mid \boldsymbol{\mu}^{(\nu)}, \boldsymbol{\pi}^{(\nu)}, y_i), \\ \pi_j^{(\nu+1)} &= w_{+j}^{(\nu+1)}, \end{aligned}$$

and let $\mu_j^{(\nu+1)}$ maximize $\sum_i w_{ij}^{(\nu)} \log(f_i(y_i|\mu_j))$. Note that the updated πs and μs maximize weighted likelihoods, and that the updated μs are not necessarily equal to the previous values. The ws are particularly straightforward to compute using Bayes' theorem:

$$w_{ij}^{(\nu+1)} \propto \pi_j^{(\nu)} f_i(y_i|\mu_j^{(\nu)}) .$$

Normalization requires dividing by the sum over index j.

For the Poisson sampling distribution, we have

$$\mu_j^{(\nu+1)} = \frac{\sum_i w_{ij}^{(\nu)} y_i}{\sum_i w_{ij}^{(\nu)}},$$

a weighted average of the y_i.

The $\mu_j^{(0)}$ should be spread out to extend somewhat beyond the range of the data. The number of mass points in the NPML is generally much smaller than k. Therefore, though the number of mass points will remain constant at J, some combination of a subset of the $\pi^{(\nu)}$ approaching, 0 and the $\mu_j^{(\nu)}$ approaching each other will occur as the EM iterations proceed. The EM algorithm has no problem dealing with this convergence to a ridge in the parameter space, though convergence will be very slow.

3.4.4 Speeding convergence and generalizations

Several authors (see, e.g., Tanner, 1993, pp. 55-57) provide methods for speeding convergence of the EM algorithm using the information decomposition (3.31) to compute an acceleration matrix. Though the approach is successful in many applications, its use moves away from the basic attraction of the EM algorithm: its ability to decompose a complicated likelihood maximization into relatively straightforward components.

Several other extensions to the EM algorithm further enhance its effectiveness. Meng and Rubin (1993) describe an algorithm designed to avoid the nested EM iterations required when the M-step is not available in closed form. Called the *Expectation/Conditional Maximization* (ECM) algorithm, it replaces the M-step with a set of conditional maximization steps. Meng and Rubin (1992) review this extension and others, including the Supplemented ECM (SECM) algorithm, which supplements the ECM algorithm to estimate the asymptotic variance matrix of the MMLE in the same way that the SEM algorithm supplements EM.

Finally, the EM algorithm also provides a starting point for Monte Carlo methods. For example, if the E-step cannot be computed analytically but $\boldsymbol{\theta}$ values can be sampled from the appropriate conditional distribution, then $\bar{S}$ can be estimated by Monte Carlo integration (see Section 5.3 for details). Wei and Tanner (1990) refer to this approach as the Monte Carlo EM (MCEM) algorithm. In fact, the entire conditional distribution of S could be estimated in this way, providing input for computing the full posterior distribution in a "data augmentation" approach (Tanner and Wong, 1987). Meng and Rubin (1992) observe that their *Partitioned ECM* (PECM) algorithm, which partitions the parameter vector into k subcomponents in order to reduce the dimensionality of the maximizations, is essentially a deterministic version of the Gibbs sampler, an extremely general and easy-to-use tool for Bayesian computation. We discuss this and other popular Markov chain Monte Carlo methods in Section 5.4.

3.5 Interval estimation

Obtaining an empirical Bayes confidence interval (EBCI) is, in principle, very simple. Given an estimated posterior $p(\theta_i|y_i, \hat{\eta})$, we could use it as we would any other posterior distribution to obtain an HPD or equal tail credible set for θ_i. That is, if $q_\alpha(y_i, \eta)$ is such that

$$P\left(\theta_i \le q_\alpha(y_i, \eta) \mid \theta_i \sim p(\theta_i|y_i, \eta)\right) = \alpha ,$$

then a $100(1-\alpha)\%$ equal tail *naive EBCI* for θ_i is

$$\left(q_{\alpha/2}(y_i, \hat{\eta}), \quad q_{1-(\alpha/2)}(y_i, \hat{\eta})\right) .$$

Why is this interval "naive?" From introductory mathematical statistics we have that

$$\text{Var}(\theta_i|\mathbf{y}) = E_{\eta|\mathbf{y}}\left[\text{Var}(\theta_i|y_i, \eta)\right] + \text{Var}_{\eta|y}\left[E(\theta_i|y_i, \eta)\right] . \qquad (3.33)$$

In the Gaussian/Gaussian case, a 95% naive EBCI would be

$$E(\theta_i \mid y_i, \hat{\eta}) \pm 1.96\sqrt{\text{Var}(\theta_i \mid y_i, \hat{\eta})} .$$

The term under the square root approximates $E_{\eta|\mathbf{y}}\left[\text{Var}(\theta_i \mid y_i, \eta)\right]$, since the EB approach simply replaces integration with maximization. But this corresponds to only the *first* term in (3.33); the naive EBCI is ignoring the posterior uncertainty about η (i.e., the second term in (3.33)). Hence, the naive interval may be too short, and have lower than advertised coverage probability.

To remedy this, we must first define the notion of "EB coverage".

Definition 3.1 $t_\alpha(\mathbf{y})$ is a $(1-\alpha)100\%$ *unconditional EB confidence set* for $g(\theta)$ if and only if for each η,

$$P_{\mathbf{y},\theta|\eta}(g(\theta) \in t_\alpha(\mathbf{y})) \approx 1 - \alpha \ .$$

So we are evaluating the performance of the EBCI over the variability inherent in *both* θ and the data. But is this too weak a requirement? Many authors have suggested instead conditioning on some data summary, b(y). For example, taking $b(\mathbf{y}) = \mathbf{y}$ produces fully conditional (Bayesian) coverage. Many likelihood theorists would take $b(\mathbf{y})$ equal to an appropriate ancillary statistic instead (see Hill, 1990). Or we might simply take $b(\mathbf{y}) = y_i$, on the grounds that y_i is sufficient for θ_i when is known.

Definition 3.2 $t_\alpha(\mathbf{y})$ is a $100(1-\alpha)\%$ *conditional EB confidence set* for $g(\theta)$ given $b(\mathbf{y})$ if, for each $b(\mathbf{y}) = b$ and η,

$$P_{\mathbf{y},\theta|b(y)=b,\eta}(g(\theta) \in t_\alpha(\mathbf{y})) \approx 1 - \alpha \ .$$

Naive EB intervals typically fail to attain their nominal coverage probability, in either the conditional or unconditional EB sense (or in the frequentist sense). One cannot usually show this analytically, but it is often easy to check via simulation. We devote the remainder of this section to outlining methods that have been proposed for "correcting" the naive EBCI.

3.5.1 Morris' approach

For the Gaussian/Gaussian model (3.9), Morris (1983a) suggests basing the EBCI on the *modified* estimated posterior that uses the "naive" mean, but inflates the variance to capture the second term in equation (3.33). This distribution is $p'(\theta_i|y_i, \hat\eta) = N(\hat\theta_i^{EB}, V^*)$, where

$$V^* = \sigma^2\left(1 - \frac{k-1}{k}\widehat{B}\right) + \frac{2}{k-3}\widehat{B}^2(Y_i - \overline{Y})^2 \ . \tag{3.34}$$

The first term in (3.34) approximates the first term in (3.33), and is essentially $Var(\theta_i|y_i, \hat\eta)$, the naive EB variance estimate. The second term in (3.34) approximates $Var_{\eta|\mathbf{y}}\left[E(\theta_i|y_i, \eta)\right]$, the second term in (3.33), and thus serves to "correct" the naive EB interval by widening it somewhat. Notice the amount of correction decreases to zero as k increases as the estimated shrinkage factor $\widehat{B}$ decreases, or as the data-value approaches the estimated prior mean.

Morris actually proposes estimating shrinkage by

$$\widehat{B} = \frac{k-3}{k-1}\left(\frac{\sigma^2}{\sigma^2 + \hat\tau^2}\right), \text{ where } \hat\tau^2 = \left[\frac{1}{k-1}\sum_i(Y_i - \overline{Y})^2 - \sigma^2\right]$$

which is the result of several "ingenious adhockeries" (see Lindley, 1983) designed to better approximate (3.33). It is important to remember that this entire derivation assumes σ^2 is known; if it is unknown, a data-based estimate may be substituted. Morris offers evidence (though not a formal proof) that his intervals do attain the desired nominal coverage. Extension of these ideas to non-Gaussian and higher dimensional settings is possible but awkward; see Morris (1988) and Christiansen and Morris (1997).

3.5.2 Marginal posterior approach

This approach is similar to Morris' in that we mimic a fully Bayesian calculation, but a bit more generally. Since η is unknown, we place a hyperprior $\psi(\eta)$ on η, and base inference about θ_i on the *marginal posterior,*

$$l_h(\theta_i|\mathbf{y}) = \int p(\theta_i|\mathbf{y},\, \eta)h(\eta|\mathbf{y})d\eta \;,$$

where $h(\eta|\mathbf{y}) \propto m(\mathbf{y}|\eta)\psi(\eta)$.

Several simplifications to this procedure are available for many models. First, if $\hat{\eta} = \hat{\eta}(\mathbf{Y})$ is sufficient for η in the marginal family $m(\mathbf{y}|\eta)$ and has density $\rho(\hat{\eta}|\eta)$, then we can replace $h(\eta|\mathbf{y})$ by $h(\eta|\hat{\eta}) \propto \rho(\hat{\eta}|\eta)\psi(\eta)$, which conditions only on a univariate statistic. Second, note that

$$\begin{aligned} p(\theta_i|\mathbf{y},\eta) &\propto p(\theta_i,\, y_i \mid y_{j\neq i},\, \eta) \\ &= p(y_i \mid \theta_i,\, y_{j\neq i},\, \eta)\, p(\theta_i \mid y_{j\neq i},\, \eta) \\ &= f(y_i \mid \theta_i,\, \eta)\, g(\theta_i \mid \eta) \;, \end{aligned} \tag{3.35}$$

the product of the sampling model and the prior. Hence $p(\theta_i \mid \mathbf{y},\, \eta)$ depends only on y_i and η, and the marginal posterior can be written as

$$l_h(\theta_i \mid \mathbf{y}) = l_h(\theta_i \mid y_i,\, \hat{\eta}) = \int p(\theta_i \mid y_i,\, \eta)h(\eta \mid \hat{\eta})d\eta \;. \tag{3.36}$$

Appropriate percentiles of the marginal posterior l_h (instead of the estimated posterior) determine the EBCI. This is an intuitively reasonable approach, since l_h accounts explicitly for the uncertainty in η. In other words, mixing $p(\theta_i \mid y,\, \eta)$ with respect to h should produce wider intervals.

As indicated by expression (3.35), the first term in the integral in equation (3.36) will typically be known due to the conjugate structure of the hierarchy. However, two issues remain. First, will h be available? And second, even if so, can the integral be computed analytically?

Deely and Lindley (1981) were the first to answer both of these questions affirmatively, obtaining closed-form results in the Poisson/gamma and "signal detection" (i.e., where θ_i is a 0-1 random variable) cases. They referred to the approach as "Bayes empirical Bayes," since placing a hyperprior on η is essentially a fully Bayesian solution to the EB problem. However, the first general method for implementing the marginal posterior approach is

$$\underbrace{\eta \to \boldsymbol{\theta} \to \mathbf{Y} \to \hat{\eta}}_{\text{data process}} \underbrace{\to \left\{ \begin{array}{c} \boldsymbol{\theta}_1^* \\ \vdots \\ \boldsymbol{\theta}_N^* \end{array} \right\} \to \left\{ \begin{array}{c} \mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_N^* \end{array} \right\} \to \left\{ \begin{array}{c} \eta_1^* \\ \vdots \\ \eta_N^* \end{array} \right\}}_{\text{bootstrap process}} .$$

Figure 3.3 *Diagram of the Laird and Louis Type III Parametric Bootstrap.*

due to Laird and Louis (1987), who used the bootstrap. Their idea was to take $h(\eta \mid \hat{\eta}) = \rho(\eta \mid \hat{\eta})$, the sampling density of $\hat{\eta}$ given η *with the arguments interchanged.* We then approximate l_ρ by observing

$$\begin{aligned} l_\rho(\theta_i \mid y_i, \hat{\eta}) &= \int p(\theta_i \mid y_i, \eta)\rho(\eta \mid \hat{\eta})d\eta \\ &\approx \frac{1}{N}\sum_{j=1}^{N} p(\theta_i \mid y_i, \eta_j^*) , \end{aligned} \tag{3.37}$$

where $\eta_j^* \overset{\text{iid}}{\sim} \rho(\cdot \mid \hat{\eta}), j = I.... N.$ Notice that this estimator converges to l_ρ as $N \to \infty$ by the Law of Large Numbers. The η_j^* values are easily generated via what the authors refer to as the *Type III Parametric Bootstrap:* given $\hat{\eta}$, we draw $\theta_i^* \sim g(\theta \mid \hat{\eta})$, and then draw $y_i^* \sim f(y \mid \theta_i^*)$, for $i = 1,...k$. From the bootstrapped data sample $\mathbf{y}^* = (y_1^*, \ldots, y_i^*)$ we may compute $\eta^* = \hat{\eta}(\mathbf{y}^*)$. Repeating this process N times, we obtain $\eta_j^* \overset{\text{iid}}{\sim} \rho(\eta \mid \hat{\eta})$, $j = 1, \ldots, N$ for use in equation (3.37), the quantiles of which, in turn, provide our corrected EBCI.

Figure 3.3 provides a. graphical illustration of the parametric bootstrap process: it simply mimics the process generating the data, replacing the unknown η by its estimate $\hat{\eta}$. The "parametric" name arises because the process is entirely generated from a single parameter estimate $\hat{\eta}$, rather than by resampling from the data vector $\mathbf{y}$ itself (a "nonparametric bootstrap"). Hence we can easily draw from $\rho(\cdot \mid \hat{\eta})$ even when this function is not available analytically (e.g., when $\hat{\eta}$ itself must be computed numerically, as in many marginal MLE settings).

Despite this computational convenience, the reader might well question the wisdom of taking $h = \rho$. An ostensibly more sensible approach would be to choose a reasonable $\psi(\eta)$, compute $h(\eta \mid \hat{\eta}) \propto \rho(\hat{\eta} \mid \eta)\psi(\eta)$, and obtain an estimator of the corresponding l_h. This marginal posterior would match a prespecified hyperprior Bayes solution, and, unlike l_ρ, would not be sensitive to the precise choice of $\hat{\eta}$. However, if *EB coverage* is the objective, then l_ρ may be as good as l_h. After all, taking quantiles of l_h will lengthen the naive EBCI, but need not "correct" the interval to level 1- α, as we desire. Still, Carlin and Gelfand (1990) show how to use the Type

III parametric bootstrap to match any l_h, *provided* the sampling density $\rho(\cdot|\hat{\eta})$ is available in closed form.

3.5.3 *Bias correction approach*

A problem with the marginal posterior approach is that it is difficult to say how to pick a *good* hyperprior $\psi(\eta)$ (i.e., one that will result in l_h quantiles that actually achieve the nominal EB coverage rate). In fact, if $\hat{\eta}(\mathbf{y})$ is badly biased, the naive EBCI may not be too short, but too *long.* Thus, in general, what is needed is not a method that will widen the naive interval, but one that will *correct* it.

Suppose we attempt to tackle the problem of EB coverage directly. Recall that $q_\alpha(y_i, \eta)$ is the α^{th} quantile of $p(\theta_i \mid y_i, \eta)$. Define

$$r(\eta. \eta. y_i, \alpha) = P(\theta_i \le q_\alpha(y_i, \hat{\eta}) \mid \theta_i \sim p(\theta_i|y_i, \eta)) \ ,$$

and

$$R(\eta, y_i, \alpha) = E_{\hat{\eta}|y_i,\eta}\left\{r(\hat{\eta}, \eta, y_i, \alpha)\right\} \ .$$

Hence R is the true EB coverage, conditional on $b(\mathbf{y}) = y_i$, of the naive EB tail area. Usually the naive EBCI is too short, i.e.,

$$R(\eta, y_i, \alpha) \begin{cases} > \alpha, & \alpha \text{ small (say .025)} \\ < \alpha, & \alpha \text{ large (say .975)} \end{cases} \ .$$

If we solved $R(\eta, y_i, \alpha') = \alpha$ for α', then using this α' with $\hat{\eta}$ would conditionally "correct the bias" in our naive procedure and give us intervals with the desired conditional EB coverage. But of course η is unknown, so instead we might solve

$$R(\hat{\eta}, y_i, \alpha') = \alpha \tag{3.38}$$

for $\alpha' = \alpha'(\hat{\eta}, y_i, \alpha)$, and take the naive interval with a replaced by α' as our corrected confidence interval. We refer to this interval as a *conditionally bias corrected* EBCI. For *unconditional* EB correction, we can replace R by

$$R^*(\eta, \alpha) = E_{\hat{\eta}, y_i|\eta}\left\{r(\hat{\eta}, \eta, y_i, \alpha)\right\} \ ,$$

and solve $R^*(\hat{\eta}, \alpha') = \alpha$ for α'. The naive interval with a replaced by this α' is called the *unconditionally bias corrected* EBCI.

Implementation

if $p(\hat{\eta} \mid y_i, \eta)$ is available in closed form (e.g., the Gaussian/Gaussian and exponential/inverse gamma models), then solving (3.38) requires only traditional numerical integration and a rootfinding algorithm. If $p(\hat{\eta} \mid y_i, \eta)$ *is not* available (e.g., if $\hat{\eta}$ itself is not available analytically, as in the Poisson/gamma and beta/binomial models), then solving (3.38) must be done via Monte Carlo methods. In particular, Carlin and Gelfand (1991a) show how the Type III parametric bootstrap may be used in this regard. Notice that in either case, since $g(\theta_i \mid \eta)$ is typically chosen to be conjugate with $f(y_i \mid \theta_i)$,

$$r(\hat{\eta}, \eta, y_i, \alpha) = F_{y_i, \eta}\left[F^{-1}_{y_i, \hat{\eta}}(\alpha)\right] ,$$

where $F_{y_i,\eta}$ is the cdf corresponding to $p(\theta_i \mid y_i, \eta)$. So even when Monte Carlo methods must be employed. mathematical subroutines can be employed at this innermost step.

To evaluate whether bias correction is truly effective, we must check whether

$$\alpha^* \equiv E_{\hat{\eta}|y_i,\eta}\left[r\left(\hat{\eta}. \eta. y_i. \alpha'(\hat{\eta}, y_i, \alpha)\right)\right] = \alpha . \quad (3.39)$$

If $\hat{\eta} \rightarrow \eta$ as $k \rightarrow \infty$ (i.e., $\hat{\eta}$ is consistent for η), then (3.39) would certainly hold for large k, but, in this event. the naive EB interval would do fine as well! For fixed k, Carlin and Gelfand (1990) provide conditions (basically, stochastic ordering of $p(\theta_i \mid y_i, \eta)$ and $p(\eta \mid y_i, \eta)$ in η) under which

$$\alpha + \min(I_1, I_2) \leq \alpha^* \leq \alpha + \max(I_1, I_2) ,$$

where

$$I_1 = \int_{\hat{\eta}>\eta} \left[\alpha'(\hat{\eta}, y_i, \alpha) - r(\hat{\eta}. \eta. y_i. \alpha'(\eta, y_i, \alpha))\right] p(\hat{\eta} \mid y_i, \alpha) d\hat{\eta} ,$$

and

$$I_2 = \int_{\hat{\eta}<\eta} \left[\alpha'(\hat{\eta}, y_i, \alpha) - r(\hat{\eta}. \eta. y_i. \alpha'(\eta, y_i, \alpha))\right] p(\hat{\eta} \mid y_i, \alpha) d\hat{\eta} .$$

That is, since $I_1 \times I_2 < 0$, the true coverage of the bias corrected EB tail area, α^*, lies in an interval containing the nominal coverage level α.

Example 3.3 We now consider the exponential/inverse gamma model, where $Y_1, \ldots, Y_k \overset{\text{iid}}{\sim} \text{Expo}(\theta_i)$ and $\theta_1, \ldots, \theta_k \overset{\text{iid}}{\sim} \text{IG}(\eta, 1)$. The marginal distribution for y_i is then $m(y_i \mid \eta) = \eta/(y_i + 1)^{\eta+1}$, $y_i > 0$, so that the MLE of η is $\hat{\eta} = k / \sum_{i=1}^{k} \log(y_i + 1)$.

Consider bias correcting the lower tail of the EBCI. Solving $R^*(\eta, \alpha') = \alpha$ for α' (unconditional bias correction), and plotting the values obtained as a function of η, we see that $\alpha'(\eta, \alpha)$ is close to a for η near 1, but decreases steadily as η increases, and that this decrease is more pronounced for larger values of α. The behavior for the upper tail is similar, except that $\alpha'(\eta, \alpha)$ is now an increasing function of η (recall $\alpha$is now near 0.95).

To evaluate the ability of the various methods considered in this section to achieve nominal EB coverage in this example, we simulated their unconditional EB coverage probabilities and average interval lengths for two nominal coverage levels (90% and 95%) where we have set the true value of $\eta = 2$ and $k = 5$. Our simulation used 3000 replications, and set $N = 400$ in those methods that required a parametric bootstrap. In addition to the methods presented in this section, we included the classical (frequentist) interval and two intervals that arise from matching a specific

Interval Method	Average Lower Endpoint	Average Upper Endpoint	Average Interval Length	Average Uncond'l Cov. Prob.
$\gamma = .90$				
Classical	.335	19.5	19.2	.901
Naive EB	.355	3.87	3.51	.839
Bias correction	.331	4.74	4.41	.897
Laird and Louis	.339	5.15	4.81	.904
ψ_1-matching	.287	3.23	2.95	.868
ψ_2-matching	.311	4.00	3.69	.894
$\gamma = .95$				
Classical	.268	39.1	38.8	.952
Naive EB	.306	5.53	5.22	.900
Bias correction	.285	7.84	7.55	.952
Laird and Louis	.283	7.79	7.50	.954
ψ_1-matching	.246	4.46	4.51	.930
ψ_2-matching	.265	5.93	5.66	.951

Table 3.4 *Comparison of simulated unconditional EB coverage probabilities, exponential/inverse gamma model.*

hyperprior Bayes solution as in equation (3.36). The two hyperpriors we consider are both noninformative and improper, namely, $\psi_1(\eta) = 1,\ \eta > 0$ (so that $h_1(\eta \mid \hat{\eta}) = G(k+1, \hat{\eta}/k)$), and $\psi_2(\eta) = 1/\eta,\ \eta > 0$ (so that $h_2(\eta \mid \hat{\eta}) = G(k, \hat{\eta}/k)$).

Looking at the results in Table 3.4, we see that the classical method faithfully produces intervals with the proper coverage level, but which are extremely long relative to all the EB methods. The naive EB intervals also perform as expected, having coverage probabilities significantly below the nominal level. The intervals based on bias correction and the Laird and Louis bootstrap approach both perform well in obtaining nominal coverage, with the former being somewhat shorter for $\gamma = .90$ and the latter a bit shorter for $\gamma = .95$. Finally, of the two hyperprior matching intervals, only the one based on ψ_2 performs well, with the ψ_1 intervals being too short. This highlights the difficulty in choosing hyperpriors that will have the desired result with respect to empirical Bayes coverage.

On a related note, the proper choice of marginal hyperparameter estimate $\hat{\eta}$ can also have an impact on coverage, and this impact is difficult to predict. Additional simulations (not shown) suggest that the best choice for $\hat{\eta}$ is not the marginal MLE, but the marginal uniformly minimum variance unbiased estimate (UMVUE), $\hat{\eta} = (k-1)/\sum_{i=1}^{k} \log(y_i + 1)$. Ironically, the MLE-based intervals were, in general, a bit too long for the bias correction method, but too short for the Laird and Louis method. ■

3.6 Generalization to regression structures

We have considered the basic two-stage empirical Bayes model with associated delta-theorem and bootstrap methods for obtaining approximately correct posterior means and variances. A brief look at the general model, where the prior mean for each coordinate can depend on a vector of regressors, shows features of this important case and further motivates the need for the more flexible and applicable hyperprior Bayesian approach to complicated models.

Consider the two-stage model where

$$\theta_i \sim N(\mathbf{X}_i\alpha, \tau^2), \tag{3.40}$$

$$Y_i|\theta_i \sim N(\theta_i, \sigma^2), \tag{3.41}$$

where $\mathbf{X}$ is a design matrix. Then, if the parameters (α, τ, σ) are known, the posterior mean and variance of θ_i are

$$E = \mathbf{X}_i\alpha + (1-B)(\mathbf{Y}_i - \mathbf{X}_i\alpha) \tag{3.42}$$

$$V = (1-B)\sigma^2, \tag{3.43}$$

with B defined as in Example 2.1. Notice that the shrinkage applies to residuals around the regression line.

For the EB approach, parameters are estimated by marginal maximum likelihood. As we have seen for the basic model, the posterior variance must be adjusted for estimation uncertainty. The Morris expansion (Morris, 1983) gives

$$\hat{V}_i = (1-\widehat{B})\sigma^2 + \frac{c}{k}\widehat{B}\sigma^2 + \tag{3.44}$$

$$+ \frac{2}{k-c-2}\widehat{B}^2(\mathbf{Y}_i - \mathbf{X}_i\hat{\alpha})^2, \tag{3.45}$$

where all hatted values are estimated and c is the dimension of a. Formula (3.44) generalizes (3.34) and is similar to the adjustment produced by using restricted maximum likelihood (REML) estimates of variance instead of ML estimates (as in SAS `Proc MIXED` or `BMDP-5V`; see Appendix C).

Derivation of (3.44) is non-trivial, and doesn't account for correlation among parameter estimates or the fact that the posterior is skewed and not Gaussian; see Laird and Louis (1987), Carlin and Gelfand (1990), and Section 3.5 of this book. As we shall see in Chapter 5, these derivations can be replaced by Monte Carlo methods that account for all of the consequences of not knowing prior parameters. The power of these methods for the linear model are further amplified for nonlinear models. For example, in

Example 3.1, to use a Gaussian/Gaussian model we transformed five-year death rates to log odds ratios. A more direct approach based on logistic regression requires the methods in Chapter 5.

3.7 Exercises

1. (Berger, 1985, p.298) Suppose $Y_i \overset{ind}{\sim} N(\theta_i, 1)$, $i = 1, \ldots, k$, and that the θ_i are i.i.d. from a common prior G. Define the marginal density $m(y_i)$ in the usual way as

$$m(y_i) = \int N(y_i \mid \theta_i, 1)\, dG(\theta_i)\ .$$

   (a) Show that, given G, the posterior mean for can be written as

$$E(\theta_i|y_i) = y_i + \frac{m'(y_i)}{m(y_i)}\ .$$

   where m' is the derivative of m.

   (b) Suggest a related nonparametric EB estimator for θ_i.

2. Under the compound sampling model (3.4) with f_i = f for all i, show that (3.5) holds (i.e.. that the Y_is are marginally i.i.d.). What is the computational significance of this result for parametric EB data analysis?

3. Consider the gamma/inverse gamma model, i.e., $Y_1, \ldots, Y_k \overset{\text{ind}}{\sim} G(\alpha, \theta_i)$, α a known tuning constant, and $\theta_1, \ldots, \theta_k \overset{\text{iid}}{\sim} IG(\eta, 1)$.

   (a) Find the marginal density of y_i, $m(y_i|\eta)$.

   (b) Suppose $\alpha = 2$. Find the marginal MLE of η, $\hat{\eta}$.

4. In the Gaussian/Gaussian model (3.9), if $\sigma^2 = 1$ and $\mu = 0$, the Y_is are marginally independent with distribution

$$Y_i|\tau \overset{iid}{\sim} N(0, 1+\tau^2) \equiv N(0, 1/B)\ .$$

   (a) Find the marginal MLE of B, $\widehat{B}_{MLE}$, and the resulting PEB point estimates $\hat{\theta}_i^{EB}$, $i = 1, \ldots, k$.

   (b) Show that while $\widehat{B}$ in (3.25) is not equal to $\widehat{B}_{MLE}$, it is unbiased for B with respect to $m(\mathbf{y}|B)$. Show further that using this $\widehat{B}$ in the usual PEB point estimation fashion produces the James-Stein estimator (3.23).

5. Prove result (3.24) using the completeness of the non-central chi-square distribution.

6. Show how to evaluate (3.24) for a general θ. (*Hint:* A non-central chi-square can be represented as a Poisson mixture of central chi-squares with mixing on the degrees of freedom.)

7. Prove result (3.26) on the maximum coordinate-specific loss for the James-Stein estimate.
8. Consider again Fisher's sleep data:

$$1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4 .$$

Suppose these k = 10 observations arose from the Gaussian/Gaussian PEB model,

$$Y_i|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma^2),\ i = 1, \ldots, k ,$$
$$\theta_i \overset{iid}{\sim} N(\mu, \tau^2),\ i = 1, \ldots, k .$$

Assuming σ^2 = 1 for these data, compute

(a) $\hat{\boldsymbol{\theta}}^{JS}$

(b) $\hat{\boldsymbol{\theta}}^{JS'}$

(c) A naive 95% EBCI for θ_9.

(d) A Morris 95% EBCI for θ_9.

Compare the two point and interval estimates you obtain.

9. Consider the PEB model

$$Y_i|\theta_i \overset{ind}{\sim} Poisson(\theta_i t_i),\ \theta_i \overset{iid}{\sim} G(a, b),\ i = 1, \ldots, k.$$

(a) Find the marginal distribution of $\mathbf{Y} = (Y_1, \ldots, Y_k)^T$.

(b) Use the method of moments to obtain closed form expressions for the hyperparameter estimates a and b. (*Hint:* Define the rates $r_i \equiv Y_i/t_i$, and equate their first two moments, $\bar{r}$ and s_r^2, to the corresponding moments in the marginal family.)

(c) Let k = 5, $a = b$ =3, and t_i = 1 for all i, and perform a simulation study to determine the actual unconditional EB coverage of the 90% equal-tail naive EBCI for θ_1.

10. Consider the data in Table 3.5. These are the numbers of pump failures, Y_i, observed in t_i thousands of hours for k = 10 different systems of a certain nuclear power plant. The observations are listed in increasing order of raw failure rate r_i, the classical point estimate of the true failure rate θ_i for the i^{th} system.

(a) Using the statistical model and results of the previous question, compute PEB point estimates and 90% equal-tail naive interval estimates for the true failure rates of systems 1, 5, 6, and 10.

(b) Using the approach of Laird and Louis (1987) and their Type III parametric bootstrap, obtain corrected 90% EBCIs for θ_5 and θ_6. Why does the θ_5 interval require more drastic correction?

11. In the Gaussian/Gaussian EM example,

i	Y_i	t_i	r_i
1	5	94.320	.053
2	1	15.720	.064
3	5	62.880	.080
4	14	125.760	.111
5	3	5.240	.573
6	19	31.440	.604
7	1	1.048	.954
8	1	1.048	.954
9	4	2.096	1.910
10	22	10.480	2.099

Table 3.5 *Pump Failure Data (Gayer and O'Muircheartaigh. 1987)*

(a) Complete the derivation of the method for finding MMLEs of the prior mean and variance.

(b) Use the Meilijson approach to find the observed information.

(c) For this model, a direct (but still iterative) approach to finding the MMLE is straightforward. Find the gradient for the marginal distribution and propose an iterative algorithm for finding the MMLEs.

(d) Find the observed information based on the marginal distribution.

(e) Compare the EM and direct approaches.

12. Do a simulation comparison of the MLE and three EB approaches: Bobbins, the Poisson/gamma, and the Gaussian/Gaussian model for estimating the rate parameter in a Poisson distribution. Note that the Gaussian/Gaussian model is incorrect. Investigate two values of k (10, 20) and several true prior distributions. Study gamma distributions with large and small means and large and small coefficients of variation. Study two distributions that are mixtures of these gammas. Compare approaches relative to SSEL and the maximum coordinate-specific loss and discuss results. You need to design the simulations.

    Extra credit: Include in your evaluation a Maritz (forced monotonicity) improvement of the Robbins rule, and the rule based on the NPML (see Section 7.1 and Subsection 3.4.3 above).

CHAPTER 4

Performance of Bayes procedures

In the preceding two chapters, we developed the Bayes and EB approaches to data analysis, and evaluated their effectiveness on their own terms (e.g., by checking the Bayes or EB coverage of the resulting confidence sets). In this chapter, we present our new technologies with a more formidable challenge: can they also emerge as effective in traditional, frequentist terms? One might expect the news to be bad here, but, for the most part, we will show that when carefully applied (usually with low-information priors), the Bayes and EB formalisms remain effective using the more traditional criteria.

Throughout this chapter, we employ the criteria outlined in Appendix B, our review of decision theory. There, using squared error loss (SEL) and its frequentist expectation, mean squared error (MSE), we develop basic examples of how shrinkage and other features of Bayes and EB procedures can offer improved performance. In this chapter, we elaborate on this idea and extend it beyond MSE performance to many other aspects of modern statistical modeling, such as the structuring of complex models and organizing of complex computations. appropriate tracking and incorporation of relevant uncertainties into a final analysis, convenient structuring of sensitivity analyses, and documentation of assumptions and models. Note that none of these features are unique to. nor place any requirements on, one's philosophical outlook.

We start with two basic, but possibly nonintuitive, examples of how Bayes rules process data. These examples serve to alert the analyst to the complex nature of multilevel inference, and drive home the point that once a model is specified, it is important to let the probability calculus take over and produce inferences. We then proceed to a frequentist and empirical Bayes evaluation of the Bayesian approach to point and interval estimation. We close with a brief discussion of experimental design, perhaps the one area where all statisticians agree on the importance of prior input.

4.1 Bayesian processing

Most of the common examples of Bayesian analysis are like Example 2.1 in that the posterior mean exhibits *shrinkage* from the data value toward the prior mean. However, univariate models with nonconjugate priors and multivariate models with conjugate priors can possibly exhibit nonintuitive behavior. In this section, we give two examples based on the Gaussian sampling distribution. Schmittlein (1989) provides additional examples.

4.1.1 Univariate stretching with a two-point prior

Consider a Gaussian sampling distribution with unit variance and a two-point prior with mass at $\pm\delta$. Based on one observation y. the posterior distribution is

$$P(\theta = \delta \mid y) = \frac{e^{2\delta y}}{1 + e^{2\delta y}},$$

with one minus this probability on $-\delta$. The posterior expectation of θ thus

$$\delta \cdot \frac{e^{2\delta y} - 1}{1 + e^{2\delta y}}\ .$$

if $\delta > 1$, for an interval of y-values around 0, the posterior mean is farther from 0 than is y, a situation we term *stretching.* Figure 4.1 illustrates the phenomenon when $\delta = 1.2$ for positive y values (the plot for negative y is a mirror image of this one). Notice that $E(\theta|y)$ stretches (exceeds y) for y less than about 1.0. For larger y, the posterior mean does exhibit the usual shrinkage back toward zero, and in fact approaches an asymptote of $E(\theta|y) = \delta$ as $y \to \infty$.

Stretching does violate intuition about what should happen with a unimodal sampling distribution. The explanation lies in use of the posterior mean (resulting from squared-error loss) in a situation where the prior is two-point. In many applications, the posterior probability itself is a sufficient summary and squared error loss is inappropriate.

4.1.2 Multivariate Gaussian model

We now consider another example where intuition may not provide the correct answer, and the Bayesian formalism is needed to sort things out. Consider again the model of Example 2.7, wherein the coordinate-specific parameters and the data are both p-variate Gaussian. That is,

$$\mathbf{Y} \mid \boldsymbol{\theta} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \quad \text{and} \quad \boldsymbol{\theta} \mid \mathbf{A} \sim N_p(\mathbf{0}, \mathbf{A})\ .$$

Writing the posterior mean and variance in their "shrinkage" form, we have

$$\begin{aligned} E(\boldsymbol{\theta} \mid \mathbf{y}) &= (\mathbf{I} - \mathbf{B})\mathbf{y} \\ Var(\boldsymbol{\theta} \mid \mathbf{y}) &= (\mathbf{I} - \mathbf{B})\mathbf{A} \end{aligned} \qquad (4.1)$$

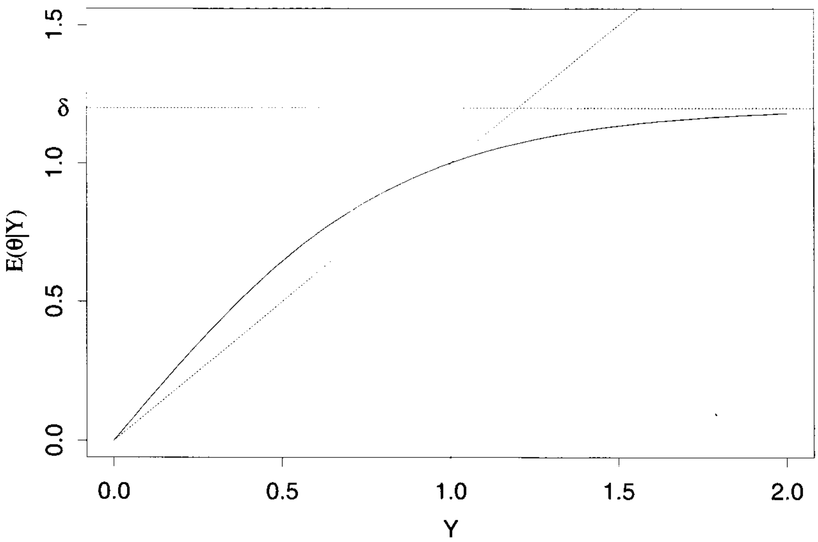


Figure 4.1 *Example of stretching: Gaussian likelihood with two-point prior.*

where $\mathbf{B} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{A})^{-1}$. If both $\boldsymbol{\Sigma}$ and $\mathbf{A}$ are diagonal, there is no linkage among coordinates, and coordinate-specific posteriors can be computed without reference to other coordinates. In general, however, the full multivariate structure must be maintained. Preserving the full structure increases efficiency (DeSouza, 1991), and can produce the nonintuitive result that coordinate-specific estimates can be stretched away from the prior mean (as in the previous example), or shrunk toward it but *beyond* it, a situation we term *crossing.* There is always shrinking toward the prior mean for the full multivariate structure. however, in that the posterior mean in (4.1) has a Euclidean norm smaller than does the data $\mathbf{y}$.

A bivariate example

Consider the bivariate case $(p = 2)$ with $\boldsymbol{\Sigma} = \mathbf{I}$ and the diagonal elements of $\mathbf{A}$ equal (i.e., the two coordinates of $\boldsymbol{\theta}$ have equal prior variance). Then $\mathbf{I}$ - $\mathbf{B}$ has the form

$$\mathbf{I} - \mathbf{B} = \begin{pmatrix} a & b \\ b & a \end{pmatrix}, \tag{4.2}$$

where $0 \le a \le 1$ and $a > |b|$. The posterior mean shows bivariate shrinkage of the observed data toward the prior mean. The point determined by the posterior mean is inside the circle with center at the prior mean, and radius the distance of the observed data from the prior mean (since I - B has eigenvalues at $\pm b$). However, the posterior mean of the first coordinate is

$$\hat{\theta}_1 = aY_1 + bY_2 \ .$$

Thus, stretching for this coordinate will occur if $Y_2 > \frac{1-a}{b}Y_1$. Crossing will occur if $Y_2 < -\frac{a}{b}Y_1$.

Though we have presented a two-stage Bayesian example. the same properties hold for empirical Bayes (see for example Waternaux et al.. 1989) and multilevel hierarchical Bayesian analyses. The general lesson is to let the Bayesian formalism operate on the full structure of the model and sort out the appropriate posterior relations, since intuitions developed in basic models may be incorrect in more complicated settings.

4.2 Frequentist performance: Point estimates

Besides having good Bayesian properties, estimators derived using Bayesian methods can have excellent frequentist and EB properties, and produce improvements over estimates generated by frequentist or likelihood-based approaches. We start with two basic examples and then a generalization. still in a basic framework. These show that the Bayesian approach can be very effective in producing attractive frequentist properties, and suggest that this performance advantage holds for more complex models.

4.2.1 Gaussian/Gaussian model

The familiar Gaussian/Gaussian model provides a simple illustration of how the Bayesian formalism can produce procedures with good frequentist properties, in many situations better than those based on maximum likelihood or unbiasedness theory. We consider the basic model wherein $Y|\theta \sim N(\theta, 1)$ and $\theta \sim N(0, \tau^2)$ (that is, the setting of Example 2.1 where $\sigma^2 = 1$ and $\mu = 0$). We compare the two decision rules $\delta_1(y) = y$ and $\delta_2(y) = (1 - B)y$, where $B = 1/(1+\tau^2)$. Here, δ_1 is the MLE and UMVUE of θ, while δ_2 is the posterior mode, median, and mean (the latter implying it is the Bayes rule under squared error loss). We then have

$$\begin{aligned} MSE(\delta_1) &= R(\theta, \delta_1) = E_{Y|\theta}(\theta - \delta_1(Y))^2 = E_{Y|\theta}(\theta - Y)^2 \\ &= Var_{Y|\theta}(Y) = 1 \ , \end{aligned}$$

and also

$$\begin{aligned} MSE(\delta_2) &= R(\theta, \delta_2) = E_{Y|\theta}(\theta - \delta_2(Y))^2 \\ &= E_{Y|\theta}[\theta - (1 - B)Y]^2 \\ &= E_{Y|\theta}[(1 - B)(\theta - Y) + B\theta]^2 \end{aligned}$$

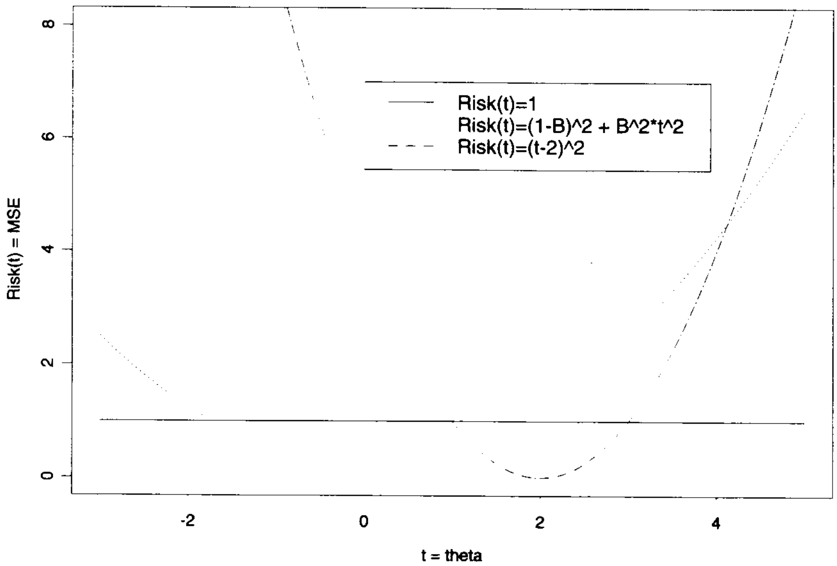


Figure 4.2 *MSE (risk under squared error loss) for three estimators in the Gaussian/Gaussian model.*

$$
\begin{aligned}
&= (1-B)^2 E_{Y|\theta}(\theta-Y)^2 \\
&\qquad\qquad +2(1-B)B\theta E_{Y|\theta}(\theta-Y) + B^2\theta^2 \\
&= (1-B)^2 + B^2\theta^2 \,,
\end{aligned}
$$

which is exactly the variance of δ_2 plus the square of its bias. It is easy to show that $MSE(\delta_2) \leq MSE(\delta_1)$ if and only if $|\theta| \leq \sqrt{(2-B)/B}$. That is, the Bayes rule has smaller *frequentist* risk provided the true mean, θ, is close to the prior mean, 0.

This situation is illustrated for the case where B = 0.5 in Figure 4.2, where the dummy argument *t is* used in place of θ. $MSE(\delta_1)$ is shown by the solid horizontal line at 1, while $MSE(\delta_2)$ is the dotted parabola centered at 0. For comparison a dashed line corresponding to the MSE for a third rule, δ_3(y) = 2, is also given. This rather silly rule always estimates θ to be 2, ignoring the data completely. Clearly its risk is given by

$$MSE(\delta_3) = R(\theta, \delta_3) = E_{Y|\theta}(\theta - 2)^2 = (\theta - 2)^2 \,,$$

which is 0 if θ happens to actually be 2, but increases much more steeply than $MSE(\delta_2)$ as θ moves away from 2. This rule is admissible, since no other rule could possibly have lower MSE for all θ (thanks to the 0

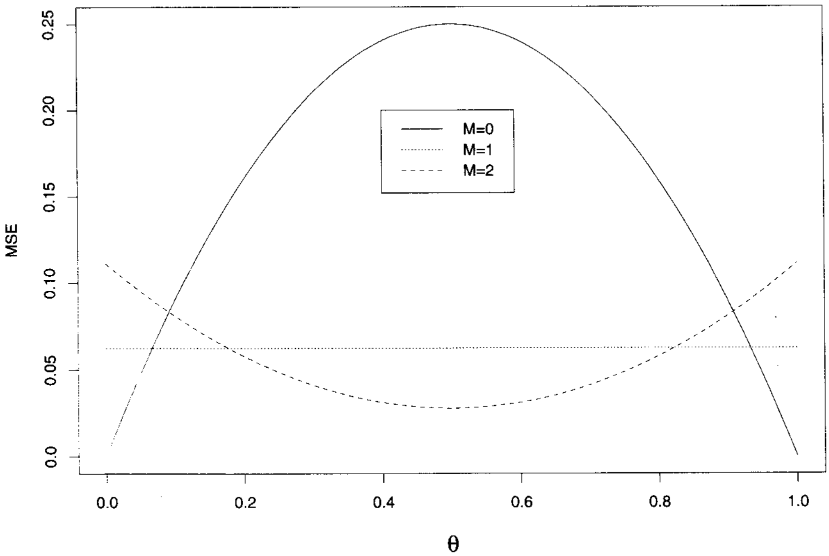


Figure 4.3 *MSE for three estimators in the beta/binomial model with n* = 1 *and* μ = 0.5.

value at δ = 2), but the large penalty paid for other θ values make it unattractive for general use. The Bayes rule δ_2 can also be shown to be admissible, but notice that it would be inadmissible if $B < 0$, since then $MSE(\delta_2) > 1 = MSE(\delta_1)$ for all θ. Fortunately, B cannot be negative since τ^2 cannot be; the point here is only to show that crossing (shrinking past the prior mean) is a poor idea in this example.

For the more general setting of Example 2.2, the Bayes rule will have smaller MSE than the MLE when

$$\theta \in \mu \pm \sqrt{\frac{\sigma^2 + 2n\tau^2}{n}} \, .$$

For the special case where $\tau^2 = \sigma^2$ (i.e., the prior is "worth" one observation), the limits simplify to $\mu \pm \sigma\sqrt{2 + \frac{1}{n}}$, a broad region of superiority.

4.2.2 Beta/binomial model

As a second example, consider again the estimation of the event probability in a binomial distribution. For the Bayesian analysis we use the conjugate *Beta(a, b)* prior distribution, and follow Subsection 3.3.2 in reparametrizing

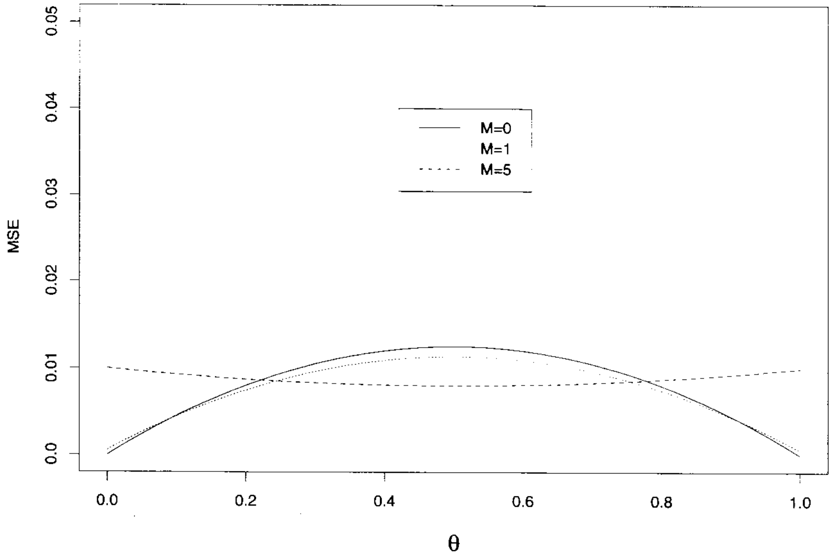


Figure 4.4 *MSE for three estimators in the beta/binomial model with* $n = 20$ *and* $\mu = 0.5$.

from (a, b) to (μ, M) where $\mu = a/(a+b)$, the prior mean, and $M = a + b$, a measure of prior precision (i.e., increasing M implies decreasing prior variance). Based on squared-error loss, the Bayes estimate is the posterior mean, namely,

$$\hat{\theta}^{\mu,M} \equiv \frac{M\mu + X}{M+n} = \frac{M}{M+n}\mu + \frac{n}{M+n}\frac{X}{n}.$$

Hence $\hat{\theta}^{\mu,M}$ is a weighted average of the prior mean and the maximum likelihood estimate, X/n. Irrespective of the value of μ, the MLE of θ is $\hat{\theta}^{\mu,0}$. The MSE of the Bayes estimate, $E_{X|\theta}(\hat{\theta}^{\mu,M} - \theta)^2$, is then given by

$$\left(\frac{n}{M+n}\right)^2 \frac{\theta(1-\theta)}{n} + \left(\frac{M}{M+n}\right)^2 (\mu-\theta)^2 . \qquad (4.4)$$

This equation shows the usual variance plus squared bias decomposition of mean squared error.

Figure 4.3 shows the risk curve for $n = 1$, $\mu = 0.5$, and $M = 0, 1, 2$. If one uses the MLE with $n = 1$, the MLE must be either 0 or l; no experienced data analyst would use such an estimator. Not surprisingly the MSE is 0 for $\theta = 0$ or 1, but it rises to .25 at $\theta = .5$. The Bayes rule

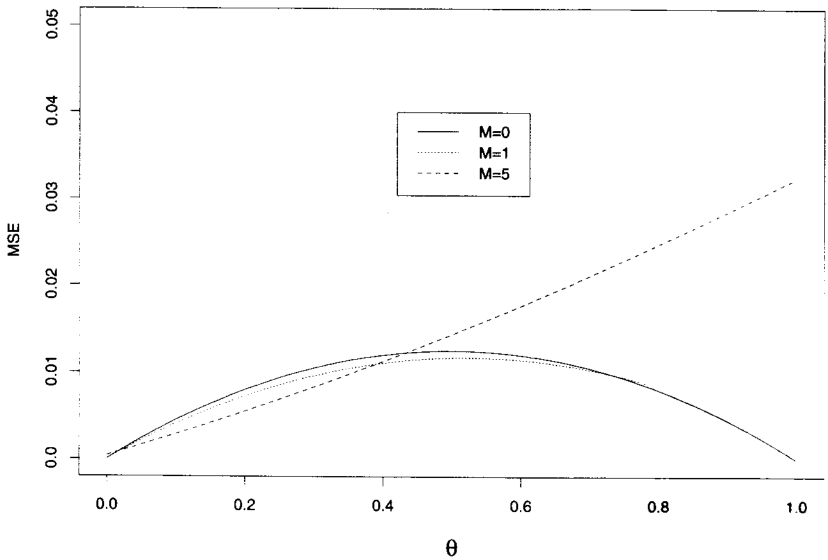


Figure 4.5 *MSE for three estimators in the beta/binomial model with n = 20 and* $\mu = 0.1$.

with $M = 1$ (dotted line) has lower MSE than the MLE for θ in the interval $.5 \pm \frac{\sqrt{3}}{4}$ = (.067-933). When $M = 2$ (dashed line), the region where the Bayes rule improves on the MLE shrinks toward 0.5, but the amount of improvement is greater. This suggests that adding a little bias to a rule in order to reduce variance can pay dividends.

Next, look at Figure 4.4, where μ is again 0.5 but now n = 20 and M = 0, 1, 5. Due to the increased sample size, all three estimators have smaller MSE than for $n = 1$, and the MLE performs quite well. The Bayes rule with $M = 1$ produces modest benefits for θ near 0.5, with little penalty for θ near 0 or 1, but it takes a larger M *(i.e.,* more weight on the prior mean) for the Bayes rule to be very different from the MLE. This is demonstrated by the curve for M = 5, which shows a benefit near 0.5, purchased by a substantial penalty for θ near 0 or 1. The data analyst (and others who need to be convinced by the analysis) would need to be quite confident that θ is near 0.5 for this estimator to be attractive.

Using the Bayes rule with "fair" prior mean μ = 0.5 and small precision $M = 1$ pays big dividends when $n = 1$ and essentially reproduces the MLE for $n = 20$. Most would agree that the MLE needs an adjustment if $n = 1$, a smaller adjustment if n = 2, and so on. Bayes estimates with diffuse priors produce big benefits for small n (where variance reduction is important).

Finally, Figure 4.5 shows the costs and benefits of using a Bayes rule with an asymmetric prior having $\mu = 0.1$ when $n = 20$. As in the symmetric prior case $\mu = 0.5$), setting 141 = 1 essentially reproduces the MLE. With M = 5, modest additional benefits accrue for θ below about 0.44 (and a little above 0), but performance is disastrous for θ greater than 0.6. Such an estimator might be attractive in some application settings, but would require near certainty that $\theta < 0.5$. We remark that the Bayes preposterior risk is the integral of these curves with respect to the prior distribution. These integrals produce the pre-posterior performance of various estimates, a subject to which we return in Section 4.4.

In summary, Bayesian point estimators based on informative priors can be risky from a frequency standpoint. showing that there is no statistical "free lunch." However, even for univariate analyses without compound sampling, the Bayesian formalism with weak prior information produces benefits for the frequentist.

A note on robustness

In both the binomial and Gaussian examples, either an empirical Bayes or hierarchical Bayes approach produces a degree of robustness to prior misspecification, at least within the parametric family, by tuning the hyperparameters to the data. In addition. for exponential sampling distributions with quadratic variance functions. Morris (1983b) has shown broad robustness outside of the conjugate prior family.

4.2.3 Generalization

The foregoing examples address two exponential family models with conjugate prior distributions. Samamego and Reneau (1994) evaluate these situations more generally for linear Bayes estimators (those where the estimate is a convex combination of the MLE and the prior mean, even if G is not conjugate) based on (iid) data when the MLE is also unbiased (and therefore MVUE). With $0 \leq \alpha < 1$ and $\hat{\theta}$ the MLE, these estimates have the form

$$\theta_G = \alpha\theta + (1-\alpha)E_G(\theta). \tag{4.5}$$

If the true prior is Go, this Bayes estimate has smaller pre-posterior MSE than the MLE so long as

$$V_{G_0}(\theta) + [E_G(\theta) - E_{G_0}(\theta)]^2 \leq \frac{1+\alpha}{1-\alpha} MSE(G_0, \hat{\theta}). \tag{4.6}$$

Since Go can be a point mass, this relation provides a standard, frequentist evaluation.

Inequality (4.6) can be explored quite generally to see when the Bayesian approach is superior. For example, if G_0 is degenerate at θ_0 then the first term is 0 and the comparison depends on the relation between θ_0, the assumed prior mean ($E_G(\theta)$) and the weight put on the prior mean (1- α This special case structures the foregoing evaluations of the beta/binomial and Gaussian/Gaussian models.

Samaniego and Reneau further investigate the situation when the true and assumed prior means are equal. Then inequality (4.6) becomes

$$V_{G_0}(\theta) \leq \frac{1+\alpha}{1-\alpha} MSE(G_0, \hat{\theta}). \tag{4.7}$$

They summarize the results of their investigation by concluding that at least for the models considered, the Bayesian approach is quite robust in that it is preferred to the frequentist so long as either the assumed prior is approximately correct or if relatively little weight is put on the prior. The Bayesian will get in trouble if she or he is overly confident about an incorrect assumption.

4.3 Frequentist performance: Confidence intervals

In addition to producing point estimates with very good frequentist performance. the Bayesian approach can also be used to develop interval estimation procedures with excellent frequentist properties in both basic and complicated modeling scenarios. As in the previous subsection, our strategy will be to use noninformative or minimally informative priors to produce the posterior distribution, which will then be used to produce intervals via either HPD or simpler "equal-tail" methods (see Subsection 2.3.2).

In complicated examples, the Bayesian formalism may be the only way to develop confidence intervals that reflect all uncertainties. We leave intervals arising from the Gaussian/Gaussian model as an exercise, and begin instead with the beta/binomial model.

4.3.1 Beta/Binomial model

Properties of the HPD interval

Again using the notation of Subsection 4.2.2, we recall that the posterior in this case is $Beta(a+x, b+n-x)$. Before considering frequentist performance, we examine how data are processed by the 100(1- 2α)% Bayesian credible interval. Suppose we adopt the uniform $Beta(a = 1, b = 1)$ prior, and consider the case where X = 0. We have the posterior distribution

$$p(\theta|X=0) = (n+1)(1-\theta)^n \ , \tag{4.8}$$

a monotone decreasing function. Thus the HPD Bayesian credible interval is one-sided. From (4.8) it is easy to show that

$$S(t|X=0) = P(\theta > t|X=0) = (1-t)^{n+1} ,$$

so letting $\tilde{\alpha} = 2\alpha$, the upper endpoint of the HPD interval may be found by solving the equation S(t) _ $\tilde{\alpha}$ for t. Doing this, we obtain the HPD interval as

$$\left(0 \, . \, 1 - \tilde{\alpha}^{\frac{1}{n+1}}\right) .$$

More generally, it is easy to show that when X = 0, the Bayesian interval is one-sided so long as $a = M\mu \leq 1$. A similar result holds for the case $X = n$, and is left as an exercise.

Unlike the case of the Gaussian/Gaussian model, however, if one uses the zero-precision prior ($M = 0$). technical problems occur if X equals either 0 or n. Hence the prior must contain at least a small amount of information. Fortunately, we have seen that using a small value of $M > 0$ actually leads to point estimates with better frequentist performance than the usual maximum likelihood estimate. We investigate the properties of the corresponding interval estimates below.

As an interesting sidelight. consider representing the upper limit in (4.9) as a pseudo number of events one might expect in a future experiment (X_n^*) divided by the sample size (n). Then.

$$X_n^* = (n+1)[1 - \tilde{\alpha}^{\frac{1}{n+1}}] \longrightarrow -\log(\tilde{\alpha}) \text{ as } n \rightarrow \infty .$$

Thus as the sample size gets large. the numerator of the upper limit of the confidence interval converges to a fixed number of pseudo-events. For example, with $\tilde{\alpha} = .05$ (i.e., $\alpha = .025$). this number of pseudo-events is - log(.05) ≈ 3. As reported by Louis (1981a) and Mann et al. (1988), this value is the upper limit for most individuals when asked: "In the light of zero events in the current experiment. what is the largest number of events you think is tenable in a future replication of the experiment?"

Frequentist performance

Based on the $Beta(a+x, b+n-x)$ posterior distribution for θ, a simple 100(1 - 2α)% Bayesian credible interval for θ is the "equal-tail" interval obtained by taking the upper and lower a-points of the posterior distribution. But as we have just seen, this approach would be silly for $a \leq 1$ when $X = 0$. This is because the monotone decreasing nature of the posterior in this case implies that the θ values excluded on the lower end would have higher posterior density than any θ values within the interval! A similar comment applies to the situation where b = (1 - μ)M $\leq$ 1 and $X = n$.

A possible remedy would be to define the "corrected equal-tail" interval as follows:

$$(\theta_L, \theta_U) = \begin{cases} (0, q_{1-\tilde{\alpha}})\ , & a + x \leq 1 \\ (q_{\tilde{\alpha}}, 1)\ , & b + n - x \leq 1 \\ (q_\alpha, q_{1-\alpha})\ , & \text{otherwise} \end{cases}\ , \tag{4.10}$$

where q_α denotes the α^{th} quantile of the $Beta(a + x, b + n - x)$ posterior, and again $\tilde{\alpha} = 2\alpha$. Though this interval is still straightforward to compute and removes the most glaring deficiency of the equal-tail procedure in this case, it will still be overly broad when the posterior is two-tailed but skewed to the left or right.

Of course, the HPD interval provides a remedy to this problem. at some computational expense. Since the one-tailed intervals in equation (4.10) are already HPD, we focus on the two-tailed case. The first step is to find the two roots θ_L and θ_U of the equation $\pi(\theta|x) = k(\alpha)$ for a given $k(\alpha)$. This is not too difficult numerically, since π in this case is simply a $(a+b+n)$-degree polynomial in θ. We next find $P(\theta_L < \theta < \theta_U|x) = F(\theta_U|x) - F(\theta_L|x)$, where F is the cdf of our beta posterior. Finally. we adjust $k(\alpha)$ up or down depending on whether the posterior coverage of our interval was too large or too small, respectively. Iterating this procedure produces the HPD interval.

Notice that the above algorithm amounts to finding the roots of the equation

$$P(\theta_L < \theta < \theta_U|x) = 1 - 2\alpha\ ,$$

wherein each step requires a rootfinder and a beta cdf routine. The latter is now readily available in many statistical software packages (such as S), but the former may require us to do some coding on our own. Fortunately, a simple *bisection* (or *interval halving)* technique will typically be easy to program and provide adequate performance. For a description of this technique, the reader is referred to a calculus or elementary numerical methods text, such as Burden et al. (1981, Section 2.1). Laird and Louis (1982) give an illustration of the method's use in an HPD interval setting.

We now examine the frequentist coverage of our Bayesian credible intervals. Adopting the notation $C(x) \equiv (\theta_L(x), \theta_U(x))$, this is defined as

$$\text{coverage}(\theta) = P_{X|\theta}[\theta \in C(X)] = E_{X|\theta}[I_{\{\theta \in C(X)\}}]\ ;$$

note carefully that θ is held fixed in this calculation. To allow for the possible effect of posterior skewness, we first compute the lower and upper non-coverages

$$\begin{aligned} \text{lower}(\theta) &= E_{X|\theta}[I_{\{\theta < \theta_L(X)\}}] \\ \text{upper}(\theta) &= E_{X|\theta}[I_{\{\theta > \theta_U(X)\}}]\ , \end{aligned} \tag{4.11}$$

so that coverage(θ = 1 - lower(θ - upper(θ). These coverages are easily

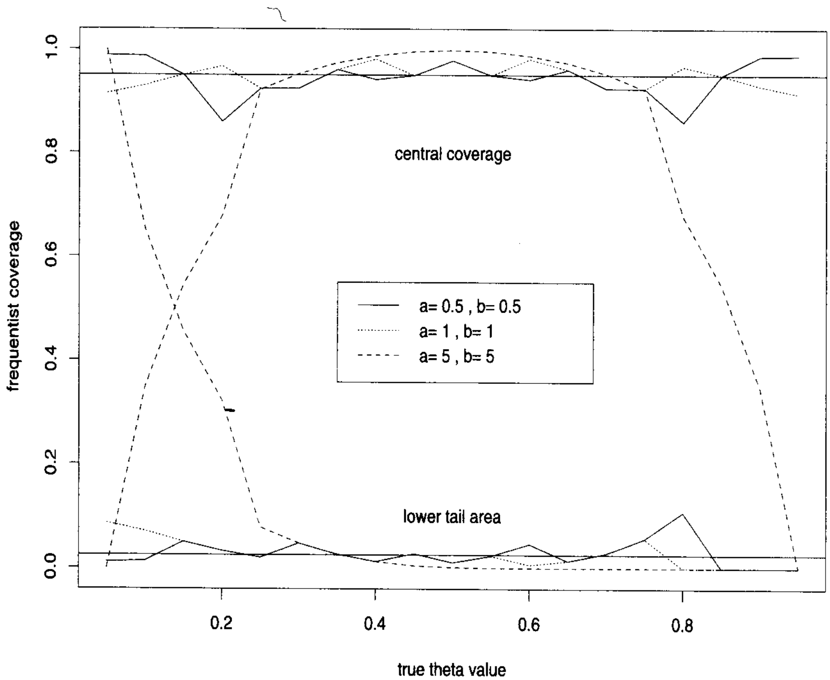


Figure 4.6 Coverage of "corrected" equal tail 95% credible interval under three symmetric priors with $n = 10$.

obtained by summing binomial probabilities. For example

$$\text{lower}(\theta) = \sum_{x=0}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} I_{\{\theta < \theta_L(x)\}}. \tag{4.12}$$

We begin by reporting coverages of the corrected equal tail intervals defined in equation (4.10) based on three beta priors that are symmetric about $\theta = 0.5$. These are the first three listed in Table 4.1, namely, the Jeffreys prior, the uniform prior, and an informative $Beta(5, 5)$ prior. The table also lists the prior mean μ, the prior precision M, and the percentage of the weight given to the prior by the posterior when $n = 10$, namely, $w_{10} = 100\frac{M}{M+10}$. Since the $Beta(5, 5)$ prior gives equal posterior weight to the data and the prior, this is clearly a strongly informative choice.

For $n = 10$, Figure 4.6 plots the coverages of the Bayesian intervals with nominal coverage probability $(1 - 2\alpha) = .95$ across a arid of possible "true" θ values from 0.05 to 0.95. Since the symmetry of the priors implies

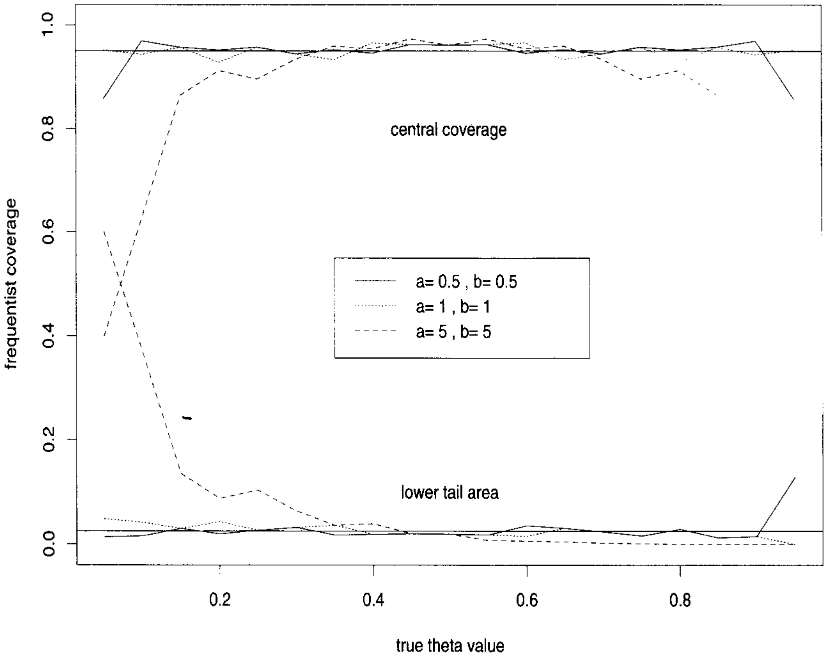


Figure 4.7 Coverage of "corrected" equal tail 95% credible interval under three symmetric priors with $n = 40$.

that lower(θ) = upper$(1 - \theta)$, the latter quantity is not shown. To enhance interpretation of the plot, horizontal reference lines are included at the 0.95 and 0.025 levels. We see that the two noninformative priors produce values for coverage(θ) that are reasonably stable at the nominal level for

a	b	μ	M	w_{10}	comments
0.5	0.5	.50	1.0	9	Jeffreys (U-shaped)
1.0	1.0	.50	2.0	17	uniform
5.0	5.0	.50	10.0	50	symmetric, strongly informative
1.0	3.0	.25	4.0	29	asymmetric, weakly informative

Table 4.1 *Beta prior distributions*

all θ, though the discrete nature of the binomial distribution causes some irregularities in the pattern. However, the informative prior performs poorly for extreme θ values, and only slightly better for θ values near 0.5 (though of course, the narrowness of its intervals should be borne in mind as well). Of particular interest is the 0% coverage provided if θ = 0.05 or 0.95, a result of the fact that the informative prior will not allow the intervals to stray this far from 0.5 - even if the data encourage such a move (i.e., X = 0 or n). As in the case of point estimation, we see that informative priors enable more precise estimation, but carry the risk of disastrous performance when their informative content is in error.

Figure 4.7 repeats the calculations of Figure 4.6 in the case where n = 40. Again both noninformative priors produce intervals with good frequentist coverage, with the uniform prior now performing noticeably better for the most extreme θ values. The informative prior also performs better, though still poorly for $|\theta - 0.5| > 0.3$. This is due to the fact that, with the increased sample size, the prior now contributes only 20% of the posterior information content, instead of 50%.

Finally, Figure 4.8 reconsiders the n = 10 case using an asymmetric Beta(1, 3) prior, listed last in Table 4.1. This prior is mildly informative in this case (contributing 4/14 = 29% of the posterior information) and favors smaller θ values over larger ones. As a result, upper(θ) is no longer a mirror image of lower(θ), and this is evident in the unacceptably large values of the former shown for large θ values. Again, the message is clear: informative priors must be used with care.

4.3.2 Fieller-Creasy problem

We now take up the problem of estimating the ratio of two normal means, a problem originally considered by Fieller (1954) and Creasy (1954). Suppose $X_1 \sim N(\theta_1, \sigma^2)$ and $X_2 \sim N(\theta_2, \sigma^2)$, assumed independent given θ_1, θ_2, and σ^2. The parameter of interest is the mean ratio $\gamma = \theta_1/\theta_2$. A plausible frequentist approach might begin by observing that $\gamma X_2 \sim N(\theta_1, \sigma^2\gamma^2)$, so that $X_1 - \gamma X_2 \sim N(0, \sigma^2(1+\gamma^2))$. Hence,

$$Pr\left(-z_{\alpha/2} \leq \frac{X_1 - \gamma X_2}{\sigma\sqrt{1+\gamma^2}} \leq z_{\alpha/2}\right) = 1 - \alpha\ ,$$

and so endpoints for a 100(1-α)% confidence interval for γ could be derived by solving the quadratic equation

$$(X_1 - \gamma X_2)^2 = z^2_{\alpha/2}\sigma^2(1+\gamma^2) \tag{4.13}$$

for γ. We notate the roots of this equation by (γ_L, γ_U).

While appearing reasonable, this approach suffers from a serious defect: the quadratic equation (4.13) may have no real roots. That is, the discrim-

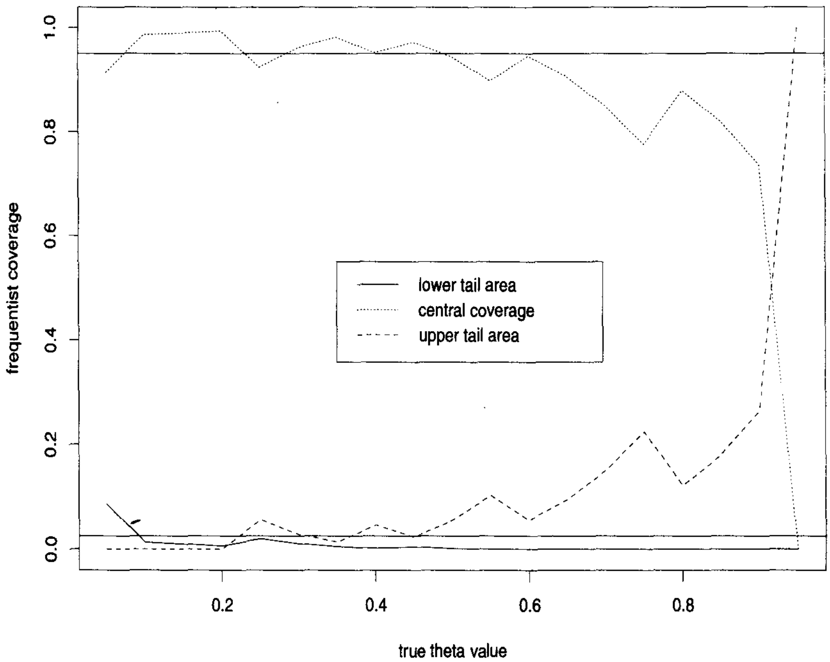


Figure 4.8 *Coverage of "corrected" equal tail 95% credible interval under an asymmetric Beta(a* $= 1$, $b = 3$*) prior with* $n = 10$.

inant of the quadratic equation,

$$D(X_1, X_2) = 4z_{\alpha/2}^2\sigma^2(X_1^2 + X_2^2 - z_{\alpha/2}^2\sigma^2) ,$$

may be negative. To remedy this, we might try the following more naive approach. Write

$$\begin{aligned}\frac{X_1}{X_2} &= \frac{X_1}{\theta_2 + X_2 - \theta_2} \\ &= \frac{X_1}{\theta_2\left(1 + \frac{X_2-\theta_2}{\theta_2}\right)} \\ &\approx \frac{X_1}{\theta_2}\left[1 - \frac{X_2 - \theta_2}{\theta_2}\right] ,\end{aligned} \tag{4.14}$$

the last line arising from a one-term Taylor expansion of the denominator on the previous line. Hence

$$\begin{aligned} E\left(\frac{X_1}{X_2}\right) &\approx \frac{1}{\theta_2} E\left[X_1 - \frac{X_1(X_2-\theta_2)}{\theta_2}\right] \\ &= \frac{1}{\theta_2}\left[\theta_1 - E(X_1)\frac{E(X_2-\theta_2)}{\theta_2}\right] \\ &= \frac{\theta_1}{\theta_2}, \end{aligned}$$

and so it makes sense to simply take $\hat{\gamma} = X_1/X_2$. Next, we may use approximation (4.14) again to obtain

$$\begin{aligned} Var\left(\frac{X_1}{X_2}\right) &\approx \frac{1}{\theta_2^2} Var\left[X_1 - \frac{X_1(X_2-\theta_2)}{\theta_2}\right] \\ &= \frac{1}{\theta_2^2}\left[\sigma^2 + Var\left(\frac{X_1(X_2-\theta_2)}{\theta_2}\right)\right. \\ &\qquad \left. -2Cov\left(X_1, \frac{X_1(X_2-\theta_2)}{\theta_2}\right)\right] \\ &= \frac{1}{\theta_2^2}\left[\sigma^2 + \frac{1}{\theta_2^2}(\sigma^2+\theta_1^2)\sigma^2 + 0\right] \\ &= \frac{\sigma^2}{\theta_2^2}\left[\frac{\sigma^2+\theta_1^2}{\theta_2^2} + 1\right], \end{aligned}$$

so a sensible variance estimate is given by

$$\widehat{Var}\left(\frac{X_1}{X_2}\right) = \frac{\sigma^2}{X_2^2}\left[\frac{\sigma^2+X_1^2}{X_2^2}+1\right].$$

Thus an alternate (1- α)100% frequentist confidence interval for γ is given by

$$\frac{X_1}{X_2} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{X_2^2}\left[\frac{\sigma^2+X_1^2}{X_2^2}+1\right]}.$$

While this procedure will never fail to produce an answer, its accuracy is certainly open to question.

By contrast, the Bayesian procedure is quite straightforward. Stephens and Smith (1992) show how to obtain samples from the joint posterior distribution of θ_1, θ_2, and σ^2 under the noninformative prior $\pi(\theta_1, \theta_2, \sigma) = 1/\sigma$ using a Monte Carlo method called the *weighted bootstrap* (see Subsection 5.3.2). We stay within the simpler σ^2 known setting, and adopt the conjugate priors $\theta_1 \sim N(\mu_1, \tau_1^2)$ and $\theta_2 \sim N(\mu_2, \tau_2^2)$ (recall that these may be made noninformative by setting $\tau_1^2 = \tau_2^2 = \infty$). From Example 2.1 we then obtain the independent posterior distributions

$$p(\theta_i|x_i) = N\left(\theta_i \,\middle|\, \frac{\sigma^2\mu_i + \tau_i^2 x_i}{\sigma^2+\tau_i^2}, \frac{\sigma^2\tau_i^2}{\sigma^2+\tau_i^2}\right), \; i = 1, 2.$$

We could now derive the distribution of $\gamma|x_1, x_2$ analytically. perhaps using the familiar Jacobian approach, but, like Stephens and Smith (1992), we find it far easier to use a sampling-based approach. Suppose we draw

$$\theta_i^{(g)} \overset{iid}{\sim} p(\theta_i|x_i),\ g = 1, \ldots, G,\ i = 1, 2,$$

and then define $\gamma^{(g)} = \theta_1^{(g)}/\theta_2^{(g)}$ g = 1, . . ., G. These $\gamma^{(g)}$ values then constitute a sample of size G from the posterior distribution of γ. Hence a (1 - α)100% equal-tail credible interval for γ could be obtained by sorting the $\{\gamma^{(g)}\}$ sample, and subsequently taking γ_L as the $(G\alpha/2)^{th}$-smallest element in the sample, and γ_U as the $(G\alpha/2)^{th}$-largest element in the sample. (Note that $(G\alpha/2)$ may first have to be rounded to the nearest integer, or else this calculation may require interpolation.)

To compare the frequentist performance of these three approaches to the Fieller-Creasy problem, we perform a simulation study. After choosing some "true" values for θ_1, θ_2, and σ^2, we generate

$$X_{ij} \overset{iid}{\sim} N(\theta_i, \sigma^2),\ j = 1, \ldots, N,\ i = 1, 2,$$

and apply each of the three procedures to each of the generated data pairs (X_{1j}, X_{2j}). (Note that to implement our sampling-based Bayesian procedure, this means we must generate $\{(\theta_{1j}^{(g)}, \theta_{2j}^{(g)}),\ g = 1,\ldots,G\}$ for each simulated data pair. Thus we effectively have a simulation within a simulation, albeit one that is very easily programmed.) For each simulated data pair, we record the lengths of the three resulting confidence intervals and whether the true γ value fell below, within, or above each interval. Clearly a method with good frequentist performance will have a coverage probability of (1 - α) or better, while at the same time producing intervals that are not too long on the average.

Table 4.2 gives the resulting average interval lengths and estimated probabilities that the true γ lies below, within, or above the computed confidence interval, where we take α = .05 and $N = G = 10{,}000$. Along with the exact and approximate frequentist approaches, we also evaluate two Bayesian methods, both of which take the two prior means μ_1 and μ_2 equal to 3. However, the first method takes $\tau_1 = \tau_2 = 1$, resulting in a highly informative prior. The second instead takes $\tau_1 = \tau_2 = 50$, a rather vague prior specification. The four portions of the table correspond to the four true (θ_1, θ_2) combinations (0,1), (0,3), (3,1), and (3,3), where $\sigma^2 = 1$ in all cases.

Looking at the table, one is first struck by the uniformly poor performance of the approximate frequentist method. While it typically produces intervals that meet or exceed the desired coverage level of 95%, they are unacceptably wide - ludicrously so when θ_2 is close to 0. The intervals based on the exact frequentist method (derived from solving the quadratic equation (4.13)) are much narrower, but in no case do they attain the nom-

	Simulated Probability			Average
	Lower	Central	Upper	Length
$(\theta_1, \theta_2) = (0, 1)$:				
Exact frequentist	0.1415	0.7151	0.1433	17.8201
(fail rate = 0.7188)				
Approx frequentist	0.0000	0.9998	0.0002	31597.4121
Informative Bayes	0.4710	0.5290	0.0000	4.8757
Vague Bayes	0.0022	0.9955	0.0023	20.5473
$(\theta_1, \theta_2) = (0, 3)$:				
Exact frequentist	0.0466	0.9082	0.0453	6.2772
(fail rate = 0.1072)				
Approx frequentist	0.0049	0.9913	0.0038	79.3618
Informative Bayes	0.5798	0.4202	0.0000	1.3461
Vague Bayes	0.0166	0.9671	0.0163	3.8475
$(\theta_1, \theta_2) = (3, 1)$:				
Exact frequentist	0.0000	0.1833	0.8167	23.1655
(fail rate = 0.0828)				
Approx frequentist	0.0000	0.8387	0.1613	5559.2930
Informative Bayes	0.0000	0.8094	0.1906	8.5788
Vague Bayes	0.0000	0.9709	0.0291	55.2689
$(\theta_1, \theta_2) = (3, 3)$:				
Exact frequentist	0.0064	0.8428	0.1508	13.5938
(fail rate = 0.0072)				
Approx frequentist	0.0000	0.9408	0.0592	25.2528
Informative Bayes	0.0016	0.9962	0.0022	1.8507
Vague Bayes	0.0024	0.9733	0.0243	8.2376

Table 4.2 *Simulation results, Fielder-Creasy problem*

inal 95% coverage level. This is no doubt attributable in part to the fact that the coverage probabilities in the table are of necessity based only on those cases where the discriminant $D(X_{1j}, X_{2j})$ was positive; the remaining cases ("failures") were simply ignored when computing the summary information in the table. With the exception of the final (θ_1, θ_2) scenario, these fail rates (shown just below the method's label in the leftmost column) are quite substantial, suggesting that whatever the method's properties, it will often be unavailable in practice.

By contrast, the performance of the two Bayesian methods is quite impressive. The intervals based on the informative prior have both very high coverage and low average length in the fourth scenario. where the prior's high confidence in $\mu_1 = \mu_2 = 3$ is justified, since these are exactly the values of θ_1 and θ_2. But naturally, this method suffers in the other cases, when the parameter values in which the method is so confident are not supported by the data. However, the Bayesian approach employing the vague prior always achieves a high level of coverage, and at a reasonable (though not overly short) average length. In fact, in two of the four cases these intervals are shorter on average than those produced by either of the frequentist methods. Thus we see that interval estimates derived from a Bayesian point of view and employing only vague (or no) prior information can be mechanically easy to compute, and also result in better frequentist performance than intervals derived from a classical approach.

4.4 Empirical Bayes performance

In Section 3.5 we showed how empirical Bayes (EB) confidence intervals may be tuned to have a desired level of *empirical Bayes* coverage, given an assumed "true" prior distribution for θ. In this section we investigate the EB coverage of Bayesian confidence sets more generally, evaluating

$$P_{Y,\theta}[\theta \in C(Y)] = E_\theta E_{Y|\theta}[I_{\{\theta \in C(Y)\}}]$$

for various modeling scenarios. We again study the Gaussian/Gaussian model, leaving others as exercises. Since EB coverage is the subject of interest, we investigate the compound sampling framework, which we recall for this problem as

$$Y_i|\theta_i \stackrel{\text{ind}}{\sim} N(\theta_i, 1),\ i = 1, \ldots, k\ , \quad \text{and} \tag{4.15}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\mu, \tau^2),\ i = 1, \ldots, k\ , \tag{4.16}$$

where for the purpose of illustration we set $k = 5$. Parallel to our development earlier in this chapter, we begin by evaluating the performance of Bayes and empirical Bayes point estimators of the θ_i in this setting, and subsequently investigate the corresponding interval estimates.

4.4.1 *Point estimation*

As discussed in Chapter 3, the traditional approach to estimation does not combine the data in the compound sampling setting. More specifically, the natural frequentist point estimate based on the likelihood (4.16) would be

$$\hat{\theta}_i^F = y_i\ ,$$

the MLE and UMVUE of θ_i. By contrast, the Bayes estimate would be the posterior mean $\hat{\theta}_i^B = B\mu + (1 - B)y_i$, where B=- 1/(1+ τ^2). Since we must specify specific values for μ and τ^2 in order to compute $\hat{\theta}_i^B$ we select the two most obvious choices $\mu = 0$ and $\tau^2 = 1$. Thus we have

$$\hat{\theta}_i^B = y_i/2 ,$$

an estimator which shrinks each observation halfway to zero. Note that this corresponds to a fairly strong prior belief, since the prior mean is receiving the same posterior weight as the data.

To avoid choosing values for μ and τ^2 we know we must either estimate them from the data or place hyperprior distributions on them. The former approach results in the empirical Bayes estimator

$$\hat{\theta}_i^{EB} = \widehat{B}\bar{y} + (1 - \widehat{B})y_i ,$$

where $\widehat{B} \equiv 1/(1+\hat{\tau}^2)$ and of course $\bar{y} = \frac{1}{k}\sum_{i=1}^k y_i$. Here we adopt the usual "positive part" estimator for τ^2 namely, $\hat{\tau}^2 = (s^2 - 1)^+ = \max\{0, s^2 - 1\}$, where $s^2 = \frac{1}{k-1}\sum_{i=1}^k (y_i - \bar{y})^2$. Alternatively, the modified EB estimate suggested by Morris (1983) takes the form

$$\hat{\theta}_i^M = \widehat{B}_M \bar{y} + (1 - \widehat{B}_M)y_i ,$$

where $\widehat{B}_M = [(k-3)/(k-1)]\widehat{B}$.

As described in Chapter 3, the Morris estimator is an attempt to approximate the hierarchical Bayes rule. But in our normal/normal setting with $k = 5$ and using the flat hyperprior $\pi(\mu, \tau^2) = 1$, Berger (1985, p.187) shows that an exact closed form expression can be obtained for the hierarchical Bayes point estimator, namely,

$$\hat{\theta}_i^{HB} = B^*\bar{y} + (1 - B^*)y_i ,$$

where

$$B^* = \frac{k-3}{(k-1)s^2}[1 - g(s^2)]$$

and

$$g(s^2) = \frac{(k-1)s^2}{2}\left\{\exp\left[\frac{(k-1)s^2}{2}\right] - 1\right\}^{-1} .$$

As in Example 3.3, we used simulation to evaluate the EB properties of our five estimators. That is, after choosing "true" values for μ and τ^2, we sampled $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{kj})$ from the prior (4.15) followed by $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{kj})$ from the likelihood (4.16), for $j = 1, \ldots, N$. This resulted in N simulated values for each of our five estimators, which we then used to estimate the EB risk under squared error loss associated with each. For example, for our final estimator we estimated the average risk

$$r = E_{\boldsymbol{\theta}} E_{\mathbf{Y}|\boldsymbol{\theta}}\left[\frac{1}{k}\sum_{i=1}^k (\hat{\theta}_i^{HB} - \theta_i)^2\right]$$

	$\hat{r}$	$\hat{se}(\hat{r})$
$(\mu, \tau^2) = (0, 1)$:		
Frequentist	1.005	0.006
Bayes ($B = 0.5$)	0.500	0.003
Empirical Bayes	0.705	0.005
Morris EB	0.750	0.005
Hierarchical Bayes	0.762	0.005
$(\mu, \tau^2) = (0, 10)$:		
Frequentist	0.994	0.006
Bayes ($B = 0.5$)	2.736	0.017
Empirical Bayes	0.976	0.006
Morris EB	0.953	0.006
Hierarchical Bayes	0.951	0.006
$(\mu, \tau^2) = (5, 1)$:		
Frequentist	1.000	0.006
Bayes ($B = 0.5$)	6.774	0.016
Empirical Bayes	0.704	0.005
Morris EB	0.746	0.005
Hierarchical Bayes	0.758	0.005
$(\mu, \tau^2) = (5, 10)$:		
Frequentist	0.987	0.006
Bayes ($B = 0.5$)	9.053	0.041
Empirical Bayes	0.975	0.006
Morris EB	0.947	0.006
Hierarchical Bayes	0.945	0.006

Table 4.3 *Simulated EB risks, Gaussian/Gaussian model with* $k = 5$

by the simulated value

$$\hat{r} = \frac{1}{Nk} \sum_{j=1}^{N} \sum_{i=1}^{K} (\hat{\theta}_{ij}^{HB} - \theta_{ij})^2$$

with associated Monte Carlo standard error estimate

$$\hat{se}(\hat{r}) = \sqrt{\frac{1}{(Nk)(Nk-1)} \sum_{j=1}^{N} \sum_{i=1}^{K} \left[(\hat{\theta}_{ij}^{HB} - \theta_{ij})^2 - \hat{r} \right]^2} .$$

Using $N = 10,000$ replications, we simulated EB risks for four possible (μ, τ^2) combinations: (0,1), (0,10), (5,1), and (5,10). The results of our simulation study are shown in Table 4.3. The fact that the frequentist estimator always has simulated risk near 1 serves as a check on our calculations, since

$$
\begin{aligned}
r(\hat{\boldsymbol{\theta}}^F) &= E_{\boldsymbol{\theta}} E_{\mathbf{Y}|\boldsymbol{\theta}} \left[\frac{1}{k} \sum_{i=1}^{k} (Y_i - \theta_i)^2 \right] \\
&= E_{\boldsymbol{\theta}} \left[\frac{1}{k} \sum_{i=1}^{k} Var_{Y_i|\theta_i}(Y_i) \right] \\
&= E_{\boldsymbol{\theta}} \left[\frac{1}{k}(k) \right] = 1 \,.
\end{aligned}
$$

The simple Bayes estimator $\hat{\boldsymbol{\theta}}^B$ performs extremely well in the first scenario (when the true μ and τ^2 are exactly those used by the estimator), but poorly when the true prior variance is inflated, and very poorly when the mean is shifted as well. As in our Section 4.2 study of frequentist risk (MSE), shrinking toward the wrong prior mean (or even overenthusiastic shrinking toward the right one) can have disastrous consequences.

By contrast. the EB performance of all three of the remaining estimators is quite impressive. being superior to that of the frequentist estimator when there is a high degree of similarity among the θ_is (i.e., τ^2 is small), and no worse otherwise. There is some indication that the hierarchical Bayes estimator does better when τ^2 is large, presumably because these values are strongly favored by the flat hyperprior assumed for this estimator. Conversely, the uncorrected EB estimator performs better when τ^2 is small. This is likely due to its tendency toward overshrinkage in the two cases where $\tau^2 = 1$, $\widehat{B}$ was observed to equal 1 (implying $\hat{\theta}^{EB} = \bar{y}$) in roughly 27% of the replications.

4.4.2 *Interval estimation*

In this subsection we continue our investigation of the Gaussian/ Gaussian compound sampling model. but in the case of evaluating the EB coverage of confidence intervals for θ_i. Once again, the frequentist procedure does no pooling of the data, resulting in the nominal 100(1 - α)% interval

$$y_i \pm z_{\alpha/2} \,,$$

where, as usual, $z_{\alpha/2}$ is the upper $(\alpha/2)$-point of the standard normal distribution.

For the Bayes interval, the symmetry of the normal posterior distribution implies that the HPD interval and equal-tail interval are equivalent in this case. Since the posterior has variance (1 - B) $=$ 0.5, we obtain the interval

$$y_i/2 \pm z_{\alpha/2}/\sqrt{2} \,,$$

which will always be only 71% as long as the frequentist interval, but will carry the risk of being inappropriately centered.

For the uncorrected ("naive") EB interval. from Section 3.5 we take the upper and lower $(\alpha/2)$-points of the *estimated* posterior, obtaining

$$\hat{\theta}_i^{EB} \pm z_{\alpha/2}\sqrt{1-\widehat{B}}\ ,$$

where $\hat{\theta}_i^{EB}$ is as given in the previous subsection. Similarly, the Morris EB interval is based on the corrected estimated posterior having mean $\hat{\theta}_i^{M}$ and variance

$$\widehat{V}_i^M = \left(1 - \frac{k-1}{k}\widehat{B}_M\right) + \frac{2}{k-3}\widehat{B}_M^2(y_i - \bar{y})^2\ ,$$

as given in Subsection 3.5.1. Thus the Morris EB interval is

$$\hat{\theta}_i^M \pm z_{\alpha/2}\sqrt{\widehat{V}_i^M}\ .$$

Finally, a flat hyperprior for (μ, τ^2) again allows the marginal posterior for θ_i (and hence. the hyperprior Bayes interval) to emerge in closed form, namely,

$$\hat{\theta}_i^{HB} \pm z_{\alpha/2}\sqrt{\widehat{V}_i^{HB}}\ ,$$

where $\hat{\theta}_i^{HB}$ is the same as before, and

$$\widehat{V}_i^{HB} = \left(1 - \frac{k-1}{k}B^*\right) + G(y_i - \bar{y})^2\ .$$

In this equation, B^* is the same as in the previous subsection, and

$$G = \frac{2(k-3)}{[(k-1)s^2]^2}\left\{1 + \left[\frac{(k-1)s^2}{2}(B^* - 1) - 1\right]g(s^2)\right\}\ .$$

where $g(s^2)$ is also the same as before.

Table 4.4 gives the results of using $N = 10,000$ replications to simulate EB coverages and interval lengths for the same four (μ, τ^2) combinations as presented in Table 4.3. The lower, central, and upper probabilities in the table correspond to the proportion of times a simulated θ_{ij} value fell below, in, or above the corresponding confidence interval. Each interval uses a nominal coverage level of 95%. That the frequentist intervals are achieving this coverage to within the accuracy of the simulation serves as a check, since the intervals are guaranteed to have the proper level of frequentist coverage and thus must also have this coverage level when averaged over any proper prior (here, a $N(\mu, \tau^2)$ distribution). Next, and as we might have guessed from the previous simulation, the Bayes interval performs very well in the first case, providing the proper coverage level and the narrowest interval of the five. However, in the other three cases (where its prior is not supported by the data), its coverage performance ranges from very poor to abysmal.

	Simulated Probability			Average
	Lower	Central	Upper	Length
$(\mu, \tau^2) = (0, 1)$:				
Frequentist	0.027	0.948	0.025	3.920
Bayes ($B = 0.5$)	0.026	0.950	0.025	2.772
Naive EB	0.186	0.629	0.185	1.964
Morris EB	0.019	0.964	0.018	3.605
Hierarchical Bayes	0.011	0.978	0.011	4.024
$(\mu, \tau^2) = (0, 10)$:				
Frequentist	0.024	0.950	0.025	3.920
Bayes ($B = 0.5$)	0.198	0.598	0.203	2.772
Naive EB	0.045	0.909	0.046	3.535
Morris EB	0.025	0.950	0.025	3.850
Hierarchical Bayes	0.021	0.957	0.022	3.972
$(\mu, \tau^2) = (5, 1)$:				
Frequentist	0.023	0.951	0.025	3.920
Bayes ($B = 0.5$)	0.000	0.057	0.943	2.772
Naive EB	0.184	0.632	0.184	1.966
Morris EB	0.017	0.965	0.018	3.605
Hierarchical Bayes	0.010	0.980	0.010	4.024
$(\mu, \tau^2) = (5, 10)$:				
Frequentist	0.024	0.952	0.025	3.920
Bayes ($B = 0.5$)	0.009	0.240	0.751	2.772
Naive EB	0.043	0.912	0.045	3.533
Morris EB	0.023	0.952	0.025	3.850
Hierarchical Bayes	0.020	0.959	0.021	3.973

Table 4.4 *EB coverages and interval lengths of nominal 95% intervals, Gaussian/Gaussian model with* $k = 5$

The naive EB interval lives up to its name in Table 4.4, missing the nominal coverage level by a small amount when τ^2 is large, and by much more when τ^2 is small. As discussed in Section 3.5, this interval is failing to capture the variability introduced by estimating μ and τ^2. But Morris' attempt at correcting this interval fares much better, providing a slightly conservative level of coverage and shorter intervals than the frequentist method - noticeably so in cases where τ^2 is small. Finally, the hierarchical Bayes method performs well when τ^2 is large, but appears overly conservative for small τ^2. Apparently if EB coverage with respect to a normal prior is the goal, a hyperprior that places too much mass on large τ^2 values will result in intervals that are a bit too long. A lighter-tailed hyperprior for τ, such as an inverse gamma, or the bias correction method for EB confidence intervals described in Section 3.5, would likely produce better results.

4.5 Design of experiments

When discussing experimental design, many of the controversies surrounding Bayesian analysis abate. Most statisticians agree that although different approaches to statistical inference engender different design goals, "everyone is a Bayesian in the design phase," even if only informally. Since no data have been collected, all evaluations are preposterior and require integration over both the data (a frequentist act) and the parameters (a Bayesian act). In the terminology of Rubin (1984), this double integration is a "Bayesianly justifiable frequentist calculation." As such, experimental design shares a kinship with the other material in this chapter, and so we include it here.

Whether the analytic goal is frequentist or Bayes, design can be structured by a likelihood, a loss function, a class of decision rules. and either formal or informal use of a prior distribution. For example, in point estimation under SEL, the frequentist may want to find a sample size sufficiently large to control the frequentist risk over a range of parameter values. The Bayesian will want a sample size sufficient to control the preposterior risk of the Bayes or other rule. Both may want procedures that are robust.

Though design is not a principal focus of this book, to show the linkage between Bayesian and frequentist approaches, we outline basic examples of Bayesian design for both frequentist and Bayesian goals. Verdinelli (1992) provides a recent review of developments in Bayesian design.

4.5.1 Bayesian design for frequentist analysis

Though everyone may be a Bayesian in the design stage. the same is most assuredly not true regarding analysis. Therefore, designs that posit a prior distribution for parameters, but design for a frequentist analysis. have an important role. We consider a basic example of finding the sample size to achieve desired type I error and power in a classical hypothesis testing model. More complicated models are addressed by Carlin and Louis (1985), Louis and Bailey (1990) and Shih (1995).

Consider one-sided hypothesis testing for the mean of n normal variables, $Y_1, \ldots, Y_n \stackrel{iid}{\sim} N(\theta, \sigma^2)$. That is, $H_0 : \theta = 0$ versus $H_a : \theta > 0$. Standard computations show that for a fixed θ, the n that achieves Type I error a and power 1 - β satisfies

$$\Phi\left(\frac{\sqrt{n}\theta}{\sigma} - z_\alpha\right) = 1 - \beta, \qquad (4.17)$$

where Φ is the standard normal cdf and, as usual, z_α is the upper α-point of this distribution. The standard frequentist design produces

$$n_\theta \quad = \quad \sigma^2 \frac{(z_\beta + z_\alpha)^2}{\theta^2} \ . \tag{4.18}$$

Since the required n depends on θ, a frequentist design will typically be based on either a typical or a conservative value of θ. However, the former can produce too small a sample size by underrepresenting uncertainty in θ, while the latter can be inefficient.

Now consider the Bayesian design for this frequentist analysis. Given a prior $g(\theta)$ for the treatment effect, Spiegelhalter (2000) suggests a simple Monte Carlo approach wherein we draw $\theta_j^* \stackrel{iid}{\sim} g(\theta_i), j = 1, \ldots, N$, and subsequently apply formula (4.17) to each outcome. This produces a collection of sample sizes $n_{\theta,j}^*$, the distribution of which reflects the uncertainty in the optimal sample size resulting from prior uncertainty. Adding a prior for σ^2 as well creates no further complication.

A slightly more formal approach would be to set the *expected* power under the chosen prior g equal to the desired level, and solve for $\tilde{n}$, i.e.,

$$\int \Phi\left(\frac{\sqrt{n}\theta}{\sigma} - z_\alpha\right) g(\theta) d\theta \quad = \quad 1 - \beta \ . \tag{4.19}$$

Shih (1995) applies a generalization of this approach to clinical trial design when the endpoint is the time to an event. Prior distributions can be placed on the treatment effect, event rates, and other inputs.

Solving for $\tilde{n}$ in (4.19) requires interval-halving or some other recursive approach. However, if G is $N(\mu, \tau^2)$, $\tilde{n}$ satisfies

$$\sqrt{\tilde{n}}\mu - z_\alpha \quad = \quad z_\beta \sqrt{1 + \tilde{n}\tau^2}. \tag{4.20}$$

For all α and β, $\tilde{n} \geq n_\theta$, and if $\tau = 0$, then $\tilde{n} = n_\theta$.

Formulas (4.17) and (4.19) produce a sample size that ensures expected power is sufficiently high. Alternatively, one could find a sample size that, for example, makes P[power > $\beta \mid n] \geq \gamma$.

Example 4.1 Consider comparison of two Poisson distributions, one with rate λ and the other with rate μ. Assume interest lies in the parameter $\delta = \lambda/\mu$ or a monotone transformation thereof. With data $S_n = X_1 + \cdots + X_n$ and $T_n = Y_1 + \cdots + Y_n$, respectively, $W_n = S_n + T_n$ is ancillary for δ (and sufficient for $\eta = \lambda + \mu$).

Irrespective of whether one is a Bayesian or frequentist, inference can be based on the conditional distribution of S_n given W_n,

$$S_n \mid W_n \quad \sim \quad Bin(W_n, \theta),$$

where $\theta = \delta/(\delta + 1)$. For this conditional model, W_n is the sample size and there is no fixed value of n that guarantees W_n will be sufficiently large.

However. Bayesian design for Bayesian analysis puts priors on both δ and η and finds a sample size that controls preposterior performance. Bayesian design for frequentist analysis puts a prior on η and finds a sample size n which produces acceptable frequentist performance in the conditional binomial distribution with parameters (W_n, θ) averaged over the distribution of W_n or otherwise using the distribution of W_n to pick a sample size. Carlin and Louis (1985) and Louis and Bailey (1990) apply these ideas to reducing problems of multiple comparisons in the analysis of rodent tumorigenicity experiments.

If sequential stopping is a design option, one can stop at $\mathcal{N}$, the first n such that W_n gives acceptable frequentist or Bayesian performance. The prior on η allows computation of the distribution of $\mathcal{N}$. ∎

4.5.2 *Bayesian design for Bayesian analysis*

Having decided that Bayesian methods are appropriate for use at the design stage, it becomes natural to consider them for the analysis stage as well. But in this case, we need to keep this new analysis plan in mind as we design the experiment. Consider for instance estimation under SEL for the model

$$Y_1, \ldots, Y_n \mid \theta \overset{iid}{\sim} f(y \mid \theta),$$

where θ has prior distribution G, and f is an exponential family with sufficient statistic $S_n = Y_1 + \cdots + Y_n$. The Bayes estimate is then $\hat{\theta}(S_n) = E_G(\theta \mid S_n)$, with posterior and preposterior risk given by

$$\rho(G, S_n) = V_G(\theta \mid S_n) \quad \text{and} \quad r(G, S_n) = E_{\mathbf{y}}[V_G(\theta \mid S_n)] \, .$$

where this last expectation is taken with respect to the marginal distribution of $\mathbf{y}$. We begin with the following simple design goal: find the smallest sample size n such that

$$r(G, S_n) \leq C^2 \, . \tag{4.21}$$

For many models, the calculations can often be accomplished in closed form, as the following two examples illustrate.

Example 4.2 Consider first the normal/normal model of Example 2.2, where both G and f are normal with variances τ^2 and σ^2, respectively. Then $\rho(G, S_n) = (1 - B_n)\sigma^2/n$, where $B_n = \sigma^2/[\sigma^2 + n\tau^2]$. This risk depends neither on the prior mean μ nor the observed data S_n. Thus $\rho(G, S_n) = r(G, S_n)$, and (4.21) produces

$$n \approx \left(\frac{\sigma}{C}\right)^2 \left[1 - \left(\frac{C}{\tau}\right)^2\right]^+ .$$

As expected, increasing C decreases n, with the opposite relation for τ. As $\tau \to \infty$, $n \to (\frac{\sigma}{C})^2$, the usual frequentist design. ∎

Example 4.3 Now consider the Poisson/gamma model of Example 2.5, where we assume that G is gamma with mean μ and variance μ^2/α, and $f(\cdot \mid \theta)$ is $Po(\theta)$. Then

$$\begin{aligned} \rho(G, S_n) &= \frac{\alpha + S_n}{(n + \alpha/\mu)^2}, \\ \text{and} \quad r(G, S_n) &= \frac{\mu^2}{\alpha + n\mu}. \end{aligned}$$

Plugging this into (4.21), we obtain

$$n \approx \frac{1}{\mu}\left[\left(\frac{\mu}{C}\right)^2 - \alpha\right]^+ .$$

For this model, the optimal Bayesian design depends on both the prior mean and variance. As $\alpha \to 0$, the prior variance increases to infinity and $n \to \mu/C^2$, which is the frequentist design for $\theta = \mu$. ■

Using preposterior Bayes risk as the design criterion produces a sufficient sample size for controlling the double integral over the sample and parameter spaces. Integrating first over the sample space represents the preposterior risk as the frequentist risk R(θ, S„.) averaged over the parameter space. The Bayesian formalism brings in uncertainty, so that generally the sample size will be larger than that obtained by a frequentist design based on a "typical" (i.e. expected) parameter value, but perhaps smaller than that based on a "conservative" value.

A fully Bayesian decision-theoretic approach to the design problem invokes the notions of expected utility and the cost of sampling. While many of the associated theoretical considerations are beyond the scope of our book, the basic ideas are not difficult to explain or understand; our discussion here follows that in Müller and Parmigiani (1995). Suppose $U(n, \mathbf{y}, \theta)$ is the payoff from an experiment having sample size n, data $\mathbf{y} = (y_1, \ldots, y_n)'$ and parameter θ, while $C(n, \mathbf{y}, \theta$ is the cost associated with this experiment. A reasonable choice for $U(n, \mathbf{y}, \theta)$ might be the negative of the loss function; i.e., under SEL we would have $U(n, \mathbf{y}, \theta) = -(\theta_n(\mathbf{y}) - \theta)^2$. Cost might be measured simply on a per sample basis, i.e., $C(n, \mathbf{y}, \theta) = cn$ for some $c > 0$. Then a plausible way of choosing the best sample size n would be to maximize the *expected utility* associated with this experiment,

$$\mathcal{U}(n) = \int \int [U(n, \mathbf{y}, \theta) - C(n, \mathbf{y}, \theta)] f(\mathbf{y}|\theta) g(\theta) d\mathbf{y} d\theta \ , \qquad (4.22)$$

where, as usual, $f(\mathbf{y}|\theta)$ is the likelihood and g(θ) the chosen prior. That is, we are choosing the design which maximizes the net payoff we expect using this sample size n, where as before we take the preposterior expectation over both the parameter θ and the (as yet unobserved) data $\mathbf{y}$. Provided the per sample cost $c > 0$, $\mathcal{U}(n)$ will typically be concave down, with expected utility at first increasing in n but eventually declining as the cost of continued sampling becomes prohibitive. Of course, since it maybe difficult to assess c in practice, one might instead simply set $c = 0$, meaning that $\mathcal{U}(n)$ would increase without bound. We could then select n as the first sample size that delivers a given utility. equivalent to the constrained minimization in condition (4.21) above when $U(n, \mathbf{y}, \theta) = -l(\theta, \hat{\theta}_n(\mathbf{y}))$.

Another difficulty is that one or both of the integrals in (4.22) may be intractable and hence require numerical integration methods. Fortunately, Monte Carlo draws are readily available by simple composition: similar to the process outlined in Figure 3.3. we would repeatedly draw θ_j^* from $g(\theta)$ followed by $\mathbf{y}_j^*$ from $f(\mathbf{y}|\theta_j^*)$. producing a joint preposterior sample $\{(\theta_j^*, \mathbf{y}_j^*), j = 1, \ldots N\}$. The associated Monte Carlo estimate of $\mathcal{U}(n)$ is

$$\widehat{\mathcal{U}}(n) = \frac{1}{N} \sum_{j=1}^{N} [U(n, \mathbf{y}_j^*, \theta_j^*) - C(n, \mathbf{y}_j^*, \theta_j^*)] \; . \tag{4.23}$$

Repeating this calculation over a grid of n values $\{n_i, \; i = 1, \ldots, I\}$, we obtain a pointwise Monte Carlo estimate of the true curve. The optimal sample size $\tilde{n}$ is the one for which $\widehat{\mathcal{U}}(n_i)$ is a max.

When exact pointwise evaluation of $\widehat{\mathcal{U}}(n_i)$ is prohibitively time consuming due to the size of I or the dimension of (y, θ), Miiller and Parmigiam (1995) recommend simply drawing a *single* (y_j^*, θ_j^*) from the joint distribution for a randomly selected design n_i, and plotting the integrands $u_i \equiv U(n, \mathbf{y}_j^*, \theta_j^*) - C(n, \mathbf{y}_j^*, \theta_j^*)$ versus n_i. Using either traditional parametric (e.g., nonlinear regression) or nonparametric (e.g.. loess smoothing) methods, we may then fit a smooth curve $\tilde{\mathcal{U}}(n)$ to the resulting point cloud, and finally take $\tilde{n} = \text{argmax}\, \tilde{\mathcal{U}}(n)$. Notice that by ignoring the integrals in (4.22) we are essentially using a Monte Carlo sample size of just $N = 1$. However, results are obtained much faster, due to the dramatic reduction in function evaluations, and may even be more accurate in higher dimensions, where the "brute force" Monte Carlo approach (4.23) may not produce a curve smooth enough for a deterministic optimization subroutine.

Example 4.4 Müller and Parmigiani (1995) consider optimal selection of the sample size for a binomial experiment. Let $f(y|\theta)$ by the usual binomial likelihood $\binom{n}{y}\theta^y(1-\theta)^{n-y}$ and $g(\theta)$ be the chosen prior - say, a *Unif* (0, 1) (a noninformative choice). If we adopt absolute error loss (AEL), an appropriate choice of payoff function is

$$U(n, \mathbf{y}, \theta) = -|m_y - \theta| \; ,$$

where m_y is the posterior median, the Bayes rule under AEL; see Appendix B, problem 3. Adopting a sampling cost of .0008 per observation, (4.22) then becomes

$$\mathcal{U}(n) = \int_0^1 \sum_{y=0}^{n} [-|m_y - \theta| - .0008n] \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta \, . \qquad (4.24)$$

Here the y sum is over a finite range, so Monte Carlo integration is needed only for θ. We thus set up a grid of n_i values from 0 to 120, and for each, draw $\theta_j^* \stackrel{iid}{\sim} Unif(0,1), j = 1, \ldots, N$, and obtain

$$\widehat{\mathcal{U}}(n_i) = \frac{1}{N} \sum_{j=0}^{N} \sum_{y=0}^{n_i} [-|m_y - \theta_j^*| - .0008n_i] \binom{n_i}{y} (\theta_j^*)^y (1-\theta_j^*)^{n_i - y} \, , \qquad (4.25)$$

where m_y can be derived numerically from the $Beta(y+1, n_i - y + 1)$ posterior. Again, $\tilde{n} = \text{argmax}\, \widehat{\mathcal{U}}(n_i)$.

Müller and Parmigiam (1995) actually note that in this case the θ integral can be done in closed form after interchanging the order of the integral and the sum in (4.24), so Monte Carlo methods are not actually needed at all. Alternatively, the quick estimate using $N = 1$ and smoothing the resulting (n_i, u_i) point cloud performs quite well in this example. ■

In summary, our intent in this chapter has been to show that procedures derived using Bayesian machinery and vague priors typically enjoy good frequentist and EB (preposterior) risk properties as well. The Bayesian approach also allows the proper propagation of uncertainty throughout the various stages of the model, and provides a convenient roadmap for obtaining estimates in complicated models. By contrast, frequentist approaches tend to rely on clever "tricks," such as sufficient statistics and pivotal quantities, that are often precluded in complex or high-dimensional models. All of this suggests the use of Bayesian methodology as an engine for producing procedures that will work well in a variety of situations.

However, there is an obvious problem with this "Bayes as procedure engine" school of thought: our ability (or lack thereof) to obtain posterior distributions in closed form. Fortunately, the recent widespread availability of high-speed computers combined with the simultaneous rapid methodological development in integration via approximate and Monte Carlo methods has provided an answer to this question. These modern computational approaches, along with guidance on their proper use and several illustrative examples, are the subject of the next chapter.

4.6 Exercises

1. In the bivariate Gaussian example of Subsection 4.1.2, show that if $\boldsymbol{\Sigma} = \mathbf{I}$, then $\mathbf{B}$ has the form (4.2).

2. In the beta/binomial point estimation setting of Subsection 4.2.2,

(a) For μ = .5 and general n and M, find the region where the Bayes rule has smaller risk than the MLE.

(b) Find the Bayes rule when the loss function is

$$L(\theta, a) = \frac{(a - \theta)^2}{\theta(1 - \theta)} .$$

3. In the beta/binomial interval estimation setting of Subsection 4.3.1,

(a) Under what condition on the prior will the HPD interval for θ be one-sided when $X = n$? Find this interval.

(b) Outline a specific interval-halving algorithm to find the HPD interval (θ_L, θ_U). Use your algorithm to find the 95% HPD interval when $a = b = 1$, $n = 5$, and $x = 1$. What percentage reduction in width does the HPD interval offer over the equal tail interval in this case?

4. Repeat the analysis plan presented for the beta/binomial in Subsection 4.3.1 for the Gaussian/Gaussian model.

(a) Show that without loss of generality you can take the prior mean (μ = 0) and the sampling variance σ^2 =1).

(b) Show that the highest posterior probability interval is also equal-tailed.

(c) Evaluate the frequentist coverage for combinations of the prior variance (τ^2) and the parameter of interest (θ).

(d) Also, evaluate the Bayesian posterior coverage and pre-posterior coverage of the interval based on μ = 0, when the true prior mean is not equal to 0.

(*Hint:* All probabilities can be represented as Gaussian integrals, and the interval endpoints come directly from the Gaussian cumulative distribution function.)

5. Show that if the sampling variance is unknown and has an inverse gamma prior distribution, then in the limit as information in this inverse gamma goes to 0 and as $\tau^2 \to \infty$, Bayesian intervals are Student's t-intervals.

6. Actually carry out the analysis in Example 4.4, where the prior $g(\theta)$ in (4.22) is

(a) the *Unif* (0,1) prior used in the example, $g(\theta) = 1$;

(b) a mixture of two *Beta* priors, $g(\theta) = .5 \cdot 3\theta^2 + .5 \cdot 30\theta^2(1 - \theta^2)$.

What is the optimal sample size $\tilde{n}$ in each case? Is either prior amenable to a fully analytical solution (i.e, without resort to Monte Carlo sampling)?

7. Using equations (4.18) and (4.20), verify that $\tilde{n} \geq n_\theta$. Will this relation hold for all G? (*Hint:* see equation (4.19)).

CHAPTER 5

Bayesian computation

5.1 Introduction

As discussed in Chapter 2, determination of posterior distributions comes down to the evaluation of complex, often high-dimensional integrals (i.e., the denominator of expression (2.1)). In addition, posterior summarization often involves computing moments or quantiles, which leads to more integration (i.e., now integrating the numerator of expression (2.1)). Some assistance was provided in Subsection 2.2.2, which described how conjugate prior forms may often be found that enable at least partial analytic evaluation of these integrals. Still, in all but the simplest model settings (typically linear models with normal likelihoods), some intractable integrations remain.

The earliest solution to this problem involved using asymptotic methods to obtain analytic approximations to the posterior density. The simplest such result is to use a normal approximation to the posterior, essentially a Bayesian version of the Central Limit Theorem. More complicated asymptotic techniques, such as Laplace's method (Tierney and Kadane, 1986), enable more accurate, possibly asymmetric posterior approximations. These methods are the subject of Subsection 5.2.

When approximate methods are intractable or result in insufficient accuracy, we must resort to numerical integration. For many years, traditional methods such as Gaussian quadrature were the methods of choice for this purpose; see Thisted (1988, Chapter 5) for a general discussion of these methods, and Naylor and Smith (1982) for their application to posterior computation. But their application was limited to models of low dimension (say, up to 10 parameters), since the necessary formulae become intractable for larger models. They also suffer from the so-called "curse of dimensionality," namely, the fact that the number of function evaluations required to maintain a given accuracy level depends exponentially on the number of dimensions. An alternative is provided by the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977), previously discussed in Section 3.4. Originally developed as a tool for finding MLEs in missing data problems, it applies equally well to finding the mode of a posterior distribution in many Bayesian settings. Indeed, much of the subsequent work broadening the scope of the algorithm to provide standard error estimates and other relevant summary information was motivated by Bayesian applications.

While the EM algorithm has seen enormous practical application. it is a fairly slow algorithm, and remains primarily geared toward finding the posterior mode, rather than estimation of the whole posterior distribution. As a result, most applied Bayesians have turned to Monte Carlo integration methods, which provide more complete information and are comparatively easy to program, even for very high-dimensional models. These methods include traditional noniterative methods such as importance sampling (Geweke, 1989) and simple rejection sampling. and are the subject of Section 5.3. Even more powerful are the more recently developed iterative Monte Carlo methods, including data augmentation (Tanner and Wong, 1987), the Metropolis-Hastings algorithm (Hastings, 1970). and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith. 1990). These methods produce a Markov chain, the output of which corresponds to a (correlated) sample from the joint posterior distribution. These methods. along with methods useful in assessing their convergence, are presented in Section 5.4. While the discussion in this final section is substantial (by far the longest in the chapter), space does not permit a full description of this enormous, rapidly growing area. A wealth of additional material may be found in the recent textbooks by Gilks et al. (1996), Gamerman (1997). Robert and Casella (1999), and Chen et al. (2000).

5.2 Asymptotic methods

5.2.1 Normal approximation

When the number of datapoints is fairly large, the likelihood will be quite peaked, and small changes in the prior will have little effect on the resulting posterior distribution. Some authors refer to this as "stable estimation." with the likelihood concentrated in a small region $\Omega \subset \Theta$ where $\pi(\theta)$ is nearly constant. In this situation, the following theorem shows that $p(\theta|\mathbf{x})$ will be approximately normal.

Theorem 5.1 Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_i(x_i|\boldsymbol{\theta})$, and thus $f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(x_i|\boldsymbol{\theta})$. Suppose the prior $\pi(\boldsymbol{\theta})$ and $f(\mathbf{x}|\boldsymbol{\theta})$ are positive. and twice differentiable near $\hat{\boldsymbol{\theta}}^{\pi}$ the posterior mode (or "generalized MLE") of $\boldsymbol{\theta}$, assumed to exist. Then under suitable regularity conditions, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ for large n can be approximated by a normal distribution having mean equal to the posterior mode, and covariance matrix equal to minus the inverse Hessian (second derivative matrix) of the log posterior evaluated at the mode. This matrix is sometimes notated as $[I^{\pi}(\mathbf{x})]^{-1}$, since it is the "generalized" observed Fisher information matrix for $\boldsymbol{\theta}$. More specifically,

$$I_{ij}^{\pi}(\mathbf{x}) = -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\left(f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\right)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{\pi}} .$$

Other forms of the normal approximation are occasionally used. For instance. if the prior is reasonably flat we might ignore it in the above calculations. This in effect replaces the posterior mode $\hat{\theta}^{\pi}$ by the MLE $\hat{\theta}$, and the generalized observed Fisher information matrix by the usual observed Fisher information matrix, $\widehat{I}(\mathbf{x})$, where

$$\begin{aligned}\widehat{I}_{ij}(x) &= -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{x}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\ &= -\sum_{l=1}^{n}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_l(x_l|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} .\end{aligned}$$

Alternatively, we might replace the posterior mode by the posterior mean $\boldsymbol{\mu}^{\pi}(\mathbf{x})$, and replace the variance estimate based on observed Fisher information with the posterior covariance matrix $V^{\pi}(\mathbf{x})$), or even the *expected* Fisher information matrix $I(\hat{\boldsymbol{\theta}})$, which in the case of lid sampling is given by

$$I_{ij}(\boldsymbol{\theta}) = -nE_{X_1|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X_1|\boldsymbol{\theta})\right] .$$

The options listed in Theorem 5.1 are the most commonly used, however, since adding the prior distribution $\pi(\boldsymbol{\theta})$ complicates matters little, and the remaining options (e.g., computing a mean instead of a mode) involve integrations - the very act the normal approximation was designed to avoid.

Just as the well-known Central Limit Theorem enables a broad range of frequentist inference by showing that the sampling distribution of the sample mean $\bar{X}$ is asymptotically normal, Theorem 5.1 enables a broad range of Bayesian inference by showing that the posterior distribution of a continuous parameter $\boldsymbol{\theta}$ is also asymptotically normal. For this reason, Theorem 5.1 is sometimes referred to as the *Bayesian Central Limit Theorem*. We do not attempt a general proof of the theorem here, but provide an outline in the unidimensional case. Write $l(\theta) = \log f(\mathbf{x}|\theta)\pi(\theta)$ and use a Taylor expansion of $l(\theta)$ as follows:

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\theta) = e^{\log f(\mathbf{x}|\theta)\pi(\theta)} = e^{l(\theta)} \\ &\approx e^{\left[l(\hat{\theta})+\frac{\partial}{\partial\theta}l(\theta)\big|_{\theta=\hat{\theta}^{\pi}}(\theta-\hat{\theta})+\frac{\partial^2}{\partial\theta^2}l(\theta)\big|_{\theta=\hat{\theta}^{\pi}}\cdot\frac{(\theta-\hat{\theta})^2}{2}\right]} \\ &= \exp\left[l(\hat{\theta})-\frac{1}{2}I^{\pi}(\mathbf{x})(\theta-\hat{\theta})^2\right] \\ &\propto N(\hat{\theta}^{\pi},\, [I^{\pi}(\mathbf{x})]^{-1}) .\end{aligned}$$

The second term in the Taylor expansion is zero, since the log posterior has slope zero at its mode, while the third term gives rise to the generalized Fisher information matrix, completing the proof. ■

Example 5.1 Consider again the beta/binomial example of Section 2.3.4. Using a flat prior on θ, we have $f(x|\theta)\pi(\theta) \propto \theta^x(1-\theta)^{n-x}$ so that $l(\theta) = x\log\theta + (n-x)\log(1-\theta)$. Taking the derivative of $l(\theta)$ and equating to zero, we obtain $\hat{\theta}^\pi = x/n$, the familiar binomial proportion. The second derivative is

$$\frac{\partial^2 l(\theta)}{\partial\theta^2} = \frac{-x}{\theta^2} - \frac{n-x}{(1-\theta)^2},$$

so that

$$\left.\frac{\partial^2 l(\theta)}{\partial\theta^2}\right|_{\theta=\hat{\theta}^\pi} = -\frac{x}{\hat{\theta}^2} - \frac{n-x}{(1-\hat{\theta})^2} = -\frac{n}{\hat{\theta}} - \frac{n}{1-\hat{\theta}}.$$

Thus $[I^\pi(x)]^{-1} = \left(\frac{n}{\hat{\theta}} + \frac{n}{1-\hat{\theta}}\right)^{-1} = \left(\frac{n}{\hat{\theta}(1-\hat{\theta})}\right)^{-1} = \frac{\hat{\theta}(1-\hat{\theta})}{n}$, which is the usual frequentist expression for the variance of $\hat{\theta} = x/n$. Thus,

$$\pi(\theta|x) \mathrel{\dot\sim} N\left(\hat{\theta}^\pi, \frac{\hat{\theta}(1-\hat{\theta})}{n}\right).$$

Figure 5.1 shows this normal approximation to the posterior of θ for the consumer preference data, previously analyzed in Section 2.3.4. Also shown is the exact posterior distribution, which we recall under the flat prior on θ is a *Beta*(14, 4) distribution. Notice that the two distributions have very similar modes, but that the approximate posterior density fails to capture the left skew in the exact posterior; it does not go to 0 at θ = 1, and drops off too quickly as θ decreases from $\hat{\theta}$ = 13/16 = .8125. This inaccuracy is also reflected in the corresponding equal tail credible sets: while the exact interval is (.57, .93), the one based on the normal distribution is (.62, 1.0). Had the observed *x* been even larger, the upper limit for the latter interval would have been greater than 1, a logical impossibility in this problem. Still, the difference is fairly subtle, especially given the small sample size involved (only 16). ■

5.2.2 Laplace's method

The posterior approximation techniques of the previous subsection are often referred to as *first order approximations* or *modal approximations*. Estimates of posterior moments and quantiles may be obtained simply as the corresponding features of the approximating normal density. But as Figure 5.1 shows, these estimates may be poor if the true posterior differs significantly from normality. This raises the question of whether we can obtain more accurate posterior estimates without significantly more effort in terms of higher order derivatives or complicated transformations.

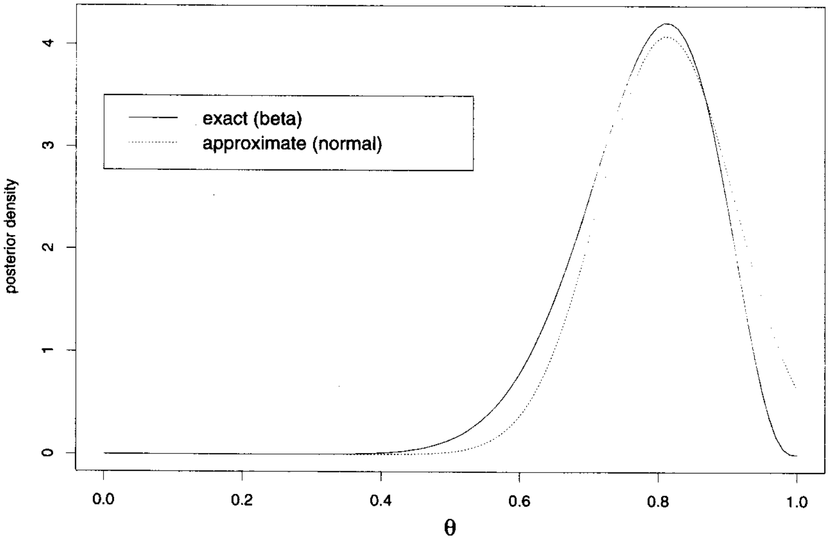


Figure 5.1 *Exact and approximate posterior distributions for* θ*, consumer preference data example.*

An affirmative answer to this question is provided by clever application of an expansion technique known as *Laplace's Method.* This technique draws its inspiration from work originally published in 1774 by Laplace (reprinted 1986); the most frequently cited paper on the subject is the rather more recent one by Tierney and Kadane (1986). The following abbreviated description follows the presentation given by Kass, Tierney, and Kadane (1988).

Suppose f is a smooth positive function of θ, and h is a smooth function of θ with $-h$ having unique maximum at $\hat{\theta}$. Again taking θ to be univariate for simplicity, we wish to approximate

$$I = \int f(\theta)e^{-nh(\theta)}d\theta \; .$$

Using Taylor series expansions of both f and h about $\hat{\theta}$, we obtain

$$\begin{aligned} I \;\; &\approx \;\; \int \left[f(\hat{\theta}) + \frac{f'(\hat{\theta})}{1!}(\theta - \hat{\theta}) + \frac{f''(\hat{\theta})}{2!}(\theta - \hat{\theta})^2 \right] \\ &\qquad \times \exp\left[-nh(\hat{\theta}) - nh'(\hat{\theta})(\theta - \hat{\theta}) - nh''(\hat{\theta})\frac{(\theta - \hat{\theta})^2}{2!} \right] d\theta \end{aligned}$$

$$\begin{aligned} &= e^{-nh(\hat{\theta})} \int \left[f(\hat{\theta}) + f'(\hat{\theta})(\theta - \hat{\theta}) + \frac{f''(\theta)}{2}(\theta - \hat{\theta})^2 \right] \\ &\quad \times \exp\left[-\frac{(\theta - \hat{\theta})^2}{2(nh''(\hat{\theta}))^{-1}} \right] d\theta \ , \end{aligned}$$

since $h'(\hat{\theta}) = 0$ by the definition of $\hat{\theta}$. But now notice that the final, multiplicative term inside the integral is proportional to a $N(\hat{\theta}, (nh''(\hat{\theta}))^{-1})$ density for θ, so that the second term inside the brackets vanishes since $E(\theta - \hat{\theta}) = 0$. Thus we have

$$I \approx e^{-nh(\hat{\theta})} f(\hat{\theta}) \sqrt{\frac{2\pi}{nh''(\hat{\theta})}} \left(1 + \frac{f''(\hat{\theta})}{2f(\hat{\theta})} \mathrm{Var}(\theta) \right) \ ,$$

where $\mathrm{Var}(\theta) = [nh''(\hat{\theta})]^{-1} = O\left(\frac{1}{n}\right)$. In general, one can show for m-dimensional $\boldsymbol{\theta}$ that $\boldsymbol{I} = \hat{I}\left\{1 + O\left(\frac{1}{n}\right)\right\}$, where

$$\hat{I} = f(\hat{\boldsymbol{\theta}}) \left(\frac{2\pi}{n} \right)^{m/2} |\tilde{\Sigma}|^{1/2} \exp\left(-nh(\hat{\boldsymbol{\theta}}) \right) \tag{5.1}$$

and $\tilde{\Sigma} = [D^2 h(\hat{\boldsymbol{\theta}})]^{-1}$. That is, $\hat{I}$ is a first order approximation to $\boldsymbol{I}$ in an asymptotic sense.

Now, suppose we wish to compute the posterior expectation of a function $g(\boldsymbol{\theta})$, where we think of $-nh(\boldsymbol{\theta})$ as the unnormalized posterior density. $\log[f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})]$. Then

$$E(g(\boldsymbol{\theta})) = \frac{\int g(\boldsymbol{\theta}) e^{-nh(\boldsymbol{\theta})} d\boldsymbol{\theta}}{\int e^{-nh(\boldsymbol{\theta})} d\boldsymbol{\theta}} \ . \tag{5.2}$$

We may apply Laplace's method (5.1) to the numerator ($f = g$) and denominator ($f = 1$) to get

$$\begin{aligned} E(g(\boldsymbol{\theta})) &= \frac{g(\hat{\boldsymbol{\theta}}) \left(\frac{2\pi}{n}\right)^{m/2} |\tilde{\Sigma}|^{1/2} \exp(-nh(\hat{\boldsymbol{\theta}})) \left[1 + O_1\left(\frac{1}{n}\right)\right]}{\left(\frac{2\pi}{n}\right)^{m/2} |\tilde{\Sigma}|^{1/2} \exp(-nh(\hat{\boldsymbol{\theta}})) \left[1 + O_2\left(\frac{1}{n}\right)\right]} \\ &= g(\hat{\boldsymbol{\theta}}) \left[\frac{1 + O_2\left(\frac{1}{n}\right) - O_2\left(\frac{1}{n}\right) + O_1\left(\frac{1}{n}\right)}{1 + O_2\left(\frac{1}{n}\right)} \right] \\ &= g(\hat{\boldsymbol{\theta}}) \left[1 + \frac{O_3\left(\frac{1}{n}\right)}{1 + O_2\left(\frac{1}{n}\right)} \right] = g(\hat{\boldsymbol{\theta}}) \left[1 + O\left(\frac{1}{n}\right) \right] \ , \end{aligned}$$

showing that $g(\hat{\boldsymbol{\theta}})$ is a first order approximation to the posterior mean of $g(\boldsymbol{\theta})$. Hence this naive application of Laplace's method produces the same estimator (with the same order of accuracy) as the modal approximation. But consider the following trick, introduced by Tierney and Kadane (1986):

provided $g(\boldsymbol{\theta}) > 0$ for all θ, write the numerator integrand in (5.2) as

$$e^{\log g(\boldsymbol{\theta}) - nh(\boldsymbol{\theta})} \equiv e^{-nh^*(\boldsymbol{\theta})} ,$$

and now use Laplace's method (5.1) with $f = 1$ in *both* the numerator and denominator. The result is

$$E(g(\boldsymbol{\theta})) = \frac{|\Sigma^*|^{1/2} \exp(-nh^*(\boldsymbol{\theta}^*))}{|\tilde{\Sigma}|^{1/2} \exp(-nh(\hat{\boldsymbol{\theta}}))} \left[1 + O\left(\frac{1}{n^2}\right)\right] , \tag{5.3}$$

where $\boldsymbol{\theta}^*$ is max of $-h^*$ and $\Sigma^* = [D^2 h^*(\boldsymbol{\theta}^*)]^{-1}$. similar to the definitions for $\hat{\boldsymbol{\theta}}$ and $\tilde{\Sigma}$ above. This clever application of Laplace's method thus provides a *second* order approximation to the posterior mean of $g(\boldsymbol{\theta})$, but still requires the calculation of only first and second derivatives of the log-posterior (although the presence of both numerator and denominator in (5.3) means we do need exactly twice as many of these calculations as we would to produce the simple normal approximation). By contrast, most other second-order expansion methods require computation of third- and perhaps higher-order derivatives. The improvement in accuracy in (5.3) comes since the leading terms in the two errors (both $O\left(\frac{1}{n}\right)$ are identical, and thus cancel when the ratio is taken (see the appendix of Tierney and Kadane, 1986, for the details of this argument).

For functions $g(\boldsymbol{\theta})$ which are not strictly positive, a simple solution is to add a large constant c to $g(\boldsymbol{\theta})$, apply Laplace's method (5.3) to this function, and finally subtract c to obtain the final answer. Alternatively, Tierney, Kass, and Kadane (1989) recommend applying Laplace's method to approximate the moment generating function $E(\exp[sg(\boldsymbol{\theta})])$, whose integrand is always positive, and then differentiating the result. These authors also show that these two approaches are asymptotically equivalent as c tends to infinity.

Improved approximation of marginal densities is also possible using this method. Suppose $\Theta = \Theta_1 \times \Theta_2$, and we want an estimate of

$$p(\theta_1|\mathbf{y}) = c \cdot \int \exp[\tilde{l}(\theta_1, \theta_2)] d\theta_2$$

where $\tilde{l}(\boldsymbol{\theta}) = \log(f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}))$ and c is an unknown normalizing constant. Taking $f = 1$ and $h = n^{-1}l$, we may use (5.3) again to get

$$\hat{p}(\theta_1|\mathbf{y}) = c^* |\tilde{\Sigma}(\theta_1)|^{1/2} \exp[\tilde{l}(\theta_1, \hat{\theta}_2(\theta_1))] , \tag{5.4}$$

where c^* is a new normalizing constant, $\hat{\theta}_2(\theta_1)$ is the maximum of $\tilde{l}(\theta_1, \cdot)$, and $\tilde{\Sigma}(\theta_1)$ is the inverse of the Hessian of $-\tilde{l}(\theta_1, \cdot)$ at $\hat{\theta}_2(\theta_1)$. In order to plot this density, formula (5.4) would need to be evaluated over a grid of θ_1 values, after which the normalizing constant c^* could be estimated. This approximation has relative error of order $O(n^{-3/2})$ on neighborhoods about the mode that shrink at the rate $n^{-1/2}$. (By contrast, the normal approximation has error of order $O(n^{-1/2})$ on these same neighborhoods.)

There are several advantages to Laplace's method. First, it is a computationally quick procedure, since it does not require an iterative algorithm. It replaces numerical integration with numerical differentiation, which is often easier and more stable numerically. It is a deterministic algorithm (i.e.. it does not rely on random numbers), so two different analysts should be able to produce a common answer given the same dataset. model, and prior distribution. Finally, it greatly reduces the computational complexity in any study of robustness, i.e., an investigation of how sensitive our conclusions are to modest changes in the prior or likelihood function. For example, suppose we wish to find the posterior expectation of a function of interest $g(\boldsymbol{\theta})$ under a new prior distribution, $\pi_{NEW}(\boldsymbol{\theta})$. We can write

$$E_{NEW}[g(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int g(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})b(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})b(\boldsymbol{\theta})d\boldsymbol{\theta}} .$$

where b(·) is the appropriate perturbation function, which in this case is $b(\boldsymbol{\theta}) = \pi_{NEW}(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})$. We can avoid starting the posterior calculations from scratch by using a result due to Kass, Tierney, and Kadane (1989), namely,

$$E_{NEW}[g(\boldsymbol{\theta})|\mathbf{y}] \approx \frac{b(\boldsymbol{\theta}^*)}{b(\hat{\boldsymbol{\theta}})} E[g(\boldsymbol{\theta})|\mathbf{y}], \tag{5.5}$$

where $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ maximize $\log[g(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})]$ and $\log[f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})]$. respectively. Note that if equation (5.3) was used to compute the original posterior expectation $E[g(\boldsymbol{\theta})|\mathbf{y}]$, then $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ are already available. Hence any number of alternate priors may be investigated simply- by evaluating the ratios $b(\boldsymbol{\theta}^*)/b(\hat{\boldsymbol{\theta}})$; no new derivative calculations are required. A somewhat more refined alternative is to obtain the original posterior expectation precisely using a more costly numerical quadrature or Monte Carlo integration method, and then use equation (5.5) to carry out a quick sensitivity analysis. We shall return to this issue in Subsection 6.1.1.

The Laplace method also has several limitations, however. For the approximation to be valid, the posterior distribution must be unimodal, or nearly so. Its accuracy also depends on the parameterization used (e.g., θ versus $\log(\theta)$), and the correct one may be difficult to ascertain. Similarly, the size of the dataset, *n*, must be fairly large; moreover, it is hard to judge how large is "large enough." That is, the second-order accuracy provided by equation (5.3) is comforting, and is surely better than the first-order accuracy provided by the normal approximation, but, for any given dataset, we will still lack a numerical measurement of how far our approximate posterior expectations are from their exact values. Worse yet, since the asymptotics are in *n*, the size of the dataset, we will not be able to improve the accuracy of our approximations without collecting additional data! Finally, for moderate- to high-dimensional $\boldsymbol{\theta}$ (say, bigger than 10), Laplace's method will rarely be of sufficient accuracy, and numerical computation of the associated Hessian matrices will be prohibitively difficult anyway. Unfortunately, high-dimensional models (such as random effects models that feature a parameter for every subject in the study) are fast becoming the norm in many branches of applied statistics.

For these reasons, practitioners now often turn to iterative methods, especially those involving Monte Carlo sampling, for posterior calculation. These methods typically involve longer runtimes, but can be applied more generally and are often easier to program. As such, they are the subject of the remaining sections in this chapter.

5.3 Noniterative Monte Carlo methods

5.3.1 Direct sampling

We begin with the most basic definition of Monte Carlo integration, found in many calculus texts. Suppose $\theta \sim h(\theta)$ and we seek $\gamma \equiv E[f(\theta)] = \int f(\theta)h(\theta)d\theta$. Then if $\theta_1, \ldots, \theta_N \overset{iid}{\sim} h(\theta)$, we have

$$\hat{\gamma} = \frac{1}{N}\sum_{j=1}^{N} f(\theta_j) \;, \tag{5.6}$$

which converges to $E[f(\theta)]$ with probability 1 as $N \to \infty$, by the Strong Law of Large Numbers. In our case, $h(\theta)$ is a posterior distribution and γ is the posterior mean of $f(\theta)$. Hence the computation of posterior expectations requires only a sample of size N from the posterior distribution.

Notice that, in contrast to the methods of Section 5.2, the quality of the approximation in (5.6) improves as we increase N, the Monte Carlo sample size (which we choose), rather than n, the size of the dataset (which is typically beyond our control). Another contrast with asymptotic methods is that the structure of (5.6) also allows us to evaluate its accuracy for any fixed N. Since $\hat{\gamma}$ is itself a sample mean of independent observations, we have that $Var(\hat{\gamma}) = Var[f(\theta)]/N$. But $Var[f(\theta)]$ can be estimated by the sample variance of the $f(\theta_i)$ values, so that a standard error estimate for $\hat{\gamma}$ is given by

$$\hat{se}(\hat{\gamma}) = \sqrt{\frac{1}{N(N-1)}\sum_{j=1}^{N}[f(\theta_j) - \hat{\gamma}]^2} \;. \tag{5.7}$$

Finally, the Central Limit Theorem implies that $\hat{\gamma} \pm 2\,\hat{se}(\hat{\gamma})$ provides an approximate 95% confidence interval for the true value of the posterior mean γ. Again, N may be chosen as large as necessary to provide any preset level of narrowness to this interval. While it may seem strange for a practical Bayesian textbook to recommend use of a frequentist interval estimation procedure. Monte Carlo simulations provide one (and perhaps the only) example where they are clearly appropriate!

In addition. letting $I_{(a,b)}$ denote the indicator function of the set (a, b), note that

$$p \equiv P\left\{a < f(\theta) < b|\mathbf{y}\right\} = E\{I_{(a,b)}[f(\theta)]|\mathbf{y}\} \ .$$

so that an estimate of p is available simply as

$$\hat{p} = \frac{\text{number of } f(\theta_j)\text{s} \in (a,b)}{N} \ ,$$

with associated binomial standard error estimate $\sqrt{\hat{p}(1-\hat{p})/N}$. In fact. this suggests that a histogram of the sampled θ_is would estimate the posterior itself, since the probability in each histogram bin converges to the true bin probability. Alternatively, we could use a kernel density estimate to "smooth" the histogram,

$$\hat{p}(\theta|\mathbf{y}) = \frac{1}{Nh_N}\sum_{j=1}^{N} K\left(\frac{\theta - \theta_j}{h_N}\right) \ ,$$

where K is a "kernel" density (typically a normal or rectangular distribution) and h_N is a window width satisfying $h_N \rightarrow 0$ and $Nh_N \rightarrow \infty$ as $N \rightarrow \infty$. The interested reader is referred to the excellent book by Silverman (1986) for more on this and other methods of density estimation.

Example 5.2 Let $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$, and suppose we adopt the reference prior $\pi(\mu, \sigma) = \frac{1}{\sigma}$. Then one can show (see Berger (1985, problem 4.12), or Lee (1997, Section 2.12)) that the joint posterior of μ and σ^2 is given by

$$\begin{aligned} \mu|\sigma^2, \mathbf{y} &\sim N(\bar{y}, \sigma^2/n) \ , \\ \text{and } \sigma^2|\mathbf{y} &\sim K\chi^{-2}_{n-1} = K \cdot IG\left(\frac{n-1}{2}, 2\right) \ , \end{aligned}$$

where $K = \sum_{i=1}^{n}(y_i - \bar{y})^2$. We may thus generate samples from the joint posterior quite easily as follows. First sample $\sigma_j^2 \sim K\chi^{-2}_{n-1}$, and then sample $\mu_j \sim N(\bar{y}, \sigma_j^2/n)$, $j = 1, \ldots, N$. This then creates the set $\{(\mu_j, \sigma_j^2),\ j = 1, \ldots, N\}$ from $p(\mu, \sigma^2|\mathbf{y})$. To estimate the posterior mean of μ, we would use

$$\hat{E}(\mu|\mathbf{y}) = \frac{1}{N}\sum_{j=1}^{N}\mu_j \ .$$

To obtain a 95% equal-tail posterior credible set for μ, we might simply use the empirical .025 and .975 quantiles of the sample of μ_j values.

Estimates of functions of the parameters are also easily obtained. For example, suppose we seek an estimate of the distribution of $\gamma = \sigma/\mu$, the coefficient of variation. We simply define the transformed Monte Carlo samples $\gamma_j = \sigma_j/\mu_j$, $j = 1, \ldots, N$, and create a histogram or kernel density estimate based on these values.

As a final illustration, suppose we wish to estimate $P(Y_{n+1} > c|\mathbf{y})$, where Y_{n+1} is a *new* observation, not part of $\mathbf{y}$. Writing $\boldsymbol{\theta} = (\mu, \sigma^2)$, we have

$$\begin{aligned} P(Y_{n+1} > c|\mathbf{y}) &= \int I_{(c,\infty)}(y_{n+1}) f(y_{n+1}|\mathbf{y}) dy_{n+1} \\ &= \int\int I_{(c,\infty)}(y_{n+1}) f(y_{n+1}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} dy_{n+1} \\ &= \int \left[\int I_{(c,\infty)}(y_{n+1}) f(y_{n+1}|\boldsymbol{\theta}) dy_{n+1}\right] p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \end{aligned}$$

the second equality coming from the fact that $f(y_{n+1}|\boldsymbol{\theta}, \mathbf{y}) = f(y_{n+1}|\boldsymbol{\theta})$, since the Y_is are conditionally independent given $\boldsymbol{\theta}$. But now the quantity inside the brackets in the third line is nothing but $P(Y_{n+1} > c|\boldsymbol{\theta}) = 1 - \Phi((c-\mu)/\sigma)$, where Φ is the cdf of a standard normal distribution. Hence

$$P(Y_{n+1} > c|\mathbf{y}) \approx \frac{1}{N}\sum_{j=1}^{N}[1 - \Phi((c-\mu_j)/\sigma_j)] \; ,$$

since the $\boldsymbol{\theta}_j \stackrel{iid}{\sim} p(\boldsymbol{\theta}|\mathbf{y})$. ■

5.3.2 Indirect methods

Example 5.2 suggests that given a sample from the posterior distribution, almost any quantity of interest can be estimated. But what if we can't directly sample from this distribution? This is an old problem that predates its interest by Bayesian statisticians by many years. As a result, there are several approaches one might try, of which we shall discuss only three: importance sampling, rejection sampling, and the weighted bootstrap.

Importance sampling

This approach is outlined carefully by Hammersley and Handscomb (1964); it has been championed for Bayesian analysis by Geweke (1989). Suppose we wish to approximate a posterior expectation, say

$$E(f(\theta)|\mathbf{y}) = \frac{\int f(\theta)L(\theta)\pi(\theta)d\theta}{\int L(\theta)\pi(\theta)d\theta} \; ,$$

where for notational convenience we again suppress any dependence of the function of interest f and the likelihood L on the data y. Suppose we can roughly approximate the normalized likelihood times prior, $cL(\theta)\pi(\theta)$, by some density $g(\theta)$ from which we can easily sample - say, a multivariate t density, or perhaps a "split-t" (i.e., a t that uses possibly different scale parameters on either side of the mode in each coordinate direction; see Geweke, 1989, for details). Then defining the *weight function* $w(\theta) = L(\theta)\pi(\theta)/g(\theta)$, we have

$$\begin{aligned} E(f(\theta)|\mathbf{y}) &= \frac{\int f(\theta)w(\theta)g(\theta)d\theta}{\int w(\theta)g(\theta)d\theta} \\ &\approx \frac{\frac{1}{N}\sum_{j=1}^{N} f(\theta_j)w(\theta_j)}{\frac{1}{N}\sum_{j=1}^{N} w(\theta_j)} , \end{aligned} \tag{5.8}$$

where $\theta_j \overset{\text{iid}}{\sim} g(\theta)$. Here, $g(\theta)$ is called the *importance function;* how closely it resembles $cL(\theta)\pi(\theta)$ controls how good the approximation in (5.8) is. To see this, note that if $g(\theta)$ is a good approximation, the weights will all be roughly equal, which in turn will minimize the variance of the numerator and denominator (see Ripley, 1987, Exercise 5.3). If on the other hand is a poor approximation, many of the weights will be close to zero, and thus a few θ_js will dominate the sums, producing an inaccurate approximation.

Example 5.3 Suppose $g(\theta)$ is taken to be the relatively light-tailed normal distribution, b but $cL(\theta)\pi(\theta)$ has much heavier, Cauchy-like tails. Then it will take many draws from g to obtain a few samples in these tails, and these points will have disproportionately large weights (since g will be small relative to $cL\pi$ for these points), thus destabilizing the estimate (5.8). As a result, a very large N will be required to obtain an approximation of acceptable accuracy. ■

We may check the accuracy of approximation (5.8) using the following formula:

$$\text{Var}\left(\frac{\bar{x}}{\bar{y}}\right) \approx \frac{\hat{\sigma}_x^2}{\bar{y}^2} + \frac{\bar{x}\hat{\sigma}_y^2}{\bar{y}^2} - \frac{\bar{x}\hat{\sigma}_{xy}}{\bar{y}^3}$$

where $\hat{\sigma}_x^2 = \frac{1}{N-1}\sum(x_j - \bar{x})^2$, $\hat{\sigma}_y^2 = \frac{1}{N-1}\sum(y_j - \bar{y})^2$, and $\hat{\sigma}_{xy} = \frac{1}{N-1}\sum(x_j - \bar{x})(y_j - \bar{y})$. (We would of course plug in $x_j = f(\theta_j)w(\theta_j)$ and $y_j = w(\theta_j)$.)

Finally, West (1992) recommends adaptive approximation of posterior densities using mixtures of multivariate t distributions. That is, after drawing a sample of size N_0 from an initial importance sampling density g_0, we compute the weighted kernel density estimate

$$g_1(\boldsymbol{\theta}) = \sum_{j=1}^{N_0} w_j K(\boldsymbol{\theta}|\boldsymbol{\theta}_j, Vh_{N_0}^2) .$$

Here, K is the density function of a multivariate t density with mode $\boldsymbol{\theta}_j$ and scale matrix $Vh_{N_0}^2$, where V is an estimate of the posterior covariance matrix and h_{N_0} is a kernel window width. We then iterate the procedure, drawing N_1 importance samples from g_1 and revising the mixture density to g_2, and so on until a suitably accurate estimate is obtained.

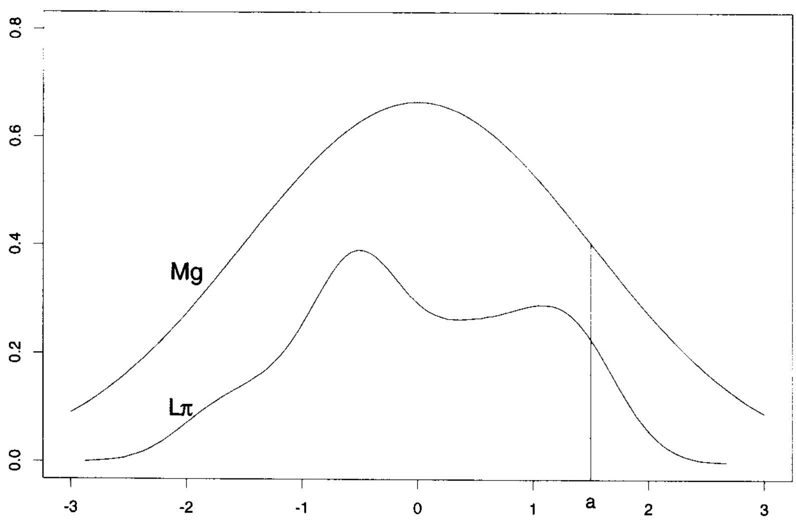


Figure 5.2 *Unstandardized posterior distribution and proper rejection envelope.*

Rejection sampling

This is an extremely general and quite common method of random generation; excellent summaries are given in the books by Ripley (1987, Section 3.2) and Devroye (1986, Section II.3). In this method, instead of trying to approximate the normalized posterior

$$h(\theta) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta} ,$$

we try to "blanket" it. That is, suppose there exists an identifiable constant $M > 0$ and a smooth density $g(\theta)$, called the *envelope function* such that $L(\theta)\pi(\theta) < Mg(\theta)$ for all θ (this situation is illustrated in Figure 5.2). The rejection method proceeds as follows:

(i) Generate $\theta_j \sim g(\theta)$.

(ii) Generate U ~ Uniform(0, 1).

(iii) If $MU\,g(\theta_j) < L(\theta_j)\pi(\theta_j)$, accept θ_j; otherwise, reject θ_j.

(iv) Return to step (i) and repeat, until the desired sample θ_j is obtained. The members of this sample will then be random variables from $h(\theta)$.

A formal proof of this result is available in Devroye (1986, pp. 40-42) or Ripley (1987. pp. 60-62); we provide only a heuristic justification. Consider a fairly large sample of points generated from $g(\theta)$. An appropriately scaled histogram of these points would have roughly the same shape as the curve labeled "Mg" in Figure 5.2. Now consider the histogram bar centered at the point labeled "a" in the figure. The rejection step in the above algorithm has the effect of slicing off the top portion of the bar (i.e.. the portion between the two curves), since only those points having -If(·g(θ_j values below the lower curve are retained. But this is true for every potential value of "a" along the horizontal axis, so a histogram of the accepted θ_j values would mimic the shape of the lower curve, which of course is proportional to the posterior distribution $h(\theta)$, as desired.

Intuition suggests that M should be chosen as small as possible, so as not to unnecessarily waste samples. This is easy to confirm, since if K denotes the number of iterations required to get one accepted candidate θ_j then K is a geometric random variable, i.e.,

$$P(K = i) = (1 - p)^{i-1} p \,, \tag{5.9}$$

where p is the probability of acceptance. So P(K = i) decreases monotonically, and at an exponential rate. It is left as an exercise to show that $p = c/M$, where c is the normalizing constant for the posterior $h(\theta)$. So our geometric distribution has mean $E(K) \quad p^{-1} = M/c$, and thus we do indeed want to minimize M. Note that if h were available for selection as the g function, we would choose the minimal acceptable value $M = c$, obtaining an acceptance probability of 1.

Like an importance sampling density, the envelope density g should be similar to the posterior in general appearance, but with heavier tails and sharper infinite peaks, in order to assure that there are sufficiently many rejection candidates available across its entire domain. One also has to be careful that Mg is actually an "envelope" for the unnormalized posterior $L\pi$. To see what happens if this condition is not met, suppose

$$S_M = \{\theta : \; L(\theta)\pi(\theta) > Mg(\theta)\} \;,$$

the situation illustrated in Figure 5.3 with S_M = (a, b). Then the distribution of the accepted θs is not $h(\theta)$, but really

$$h^*(\theta) = \begin{cases} \dfrac{L(\theta)\pi(\theta)}{\int_{S_M^C} L(\theta)\pi(\theta)d\theta + MP_g(S_M)} \,, & \theta \in S_M^C \\ \dfrac{Mg(\theta)}{\int_{S_M^C} L(\theta)\pi(\theta)d\theta + MP_g(S_M)} \,, & \theta \in S_M \end{cases} \tag{5.10}$$

Unfortunately, $\int_{S_M} L(\theta)\pi(\theta)d\theta > MP_g(S_M)$, so even if $P_g(S_M)$ is small, there is no guarantee that (5.10) is close to $h(\theta)$. That is, only a few observed envelope violations do not necessarily imply a small inaccuracy in the posterior sample.

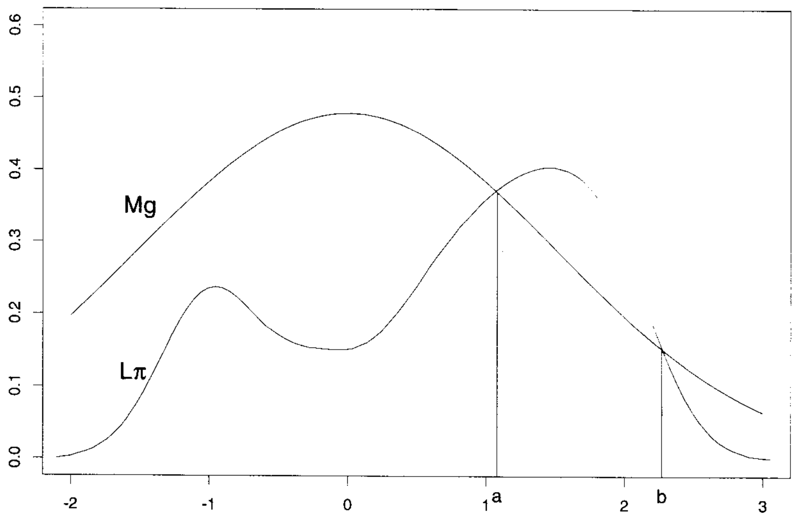


Figure 5.3 Unstandardized posterior distribution and deficient rejection envelope.

As a possible solution, when we find a θ_j such that $L(\theta_j)\pi(\theta_j) > Mg(\theta_j)$, we may do a local search in the neighborhood of θ_j and increase M accordingly. Of course, we should really go back and recheck all of the previously accepted θ_js, since some may no longer be acceptable with the new, larger M. We discuss rejection algorithms designed to eliminate envelope violations following Example 5.6 below.

Weighted bootstrap

This method was presented by Smith and Gelfand (1992), and is very similar to the *sampling-importance resampling* algorithm of Rubin (1988). Suppose an M appropriate for the rejection method is not readily available, but that we do have a sample $\theta_1, \ldots \theta_N$ from some approximating density $g(\theta)$. Define

$$w_i = \frac{L(\theta_i)\pi(\theta_i)}{g(\theta_i)} \quad \text{and} \quad q_i = \frac{w_i}{\sum_{j=1}^{N} w_i} .$$

Now draw θ^* from the *discrete* distribution over $\{\theta_1 \ldots \theta_N\}$ which places mass qi at θ_i. Then

$$\theta^* \dot{\sim} h(\theta) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta} ,$$

with the approximation improving as $N \to \infty$. This is a *weighted* bootstrap, since instead of resampling from the set $\{\theta_1, \ldots \theta_N\}$ with equally likely probabilities of selection, we are resampling some points more often than others due to the unequal weighting.

To see that the method does perform as advertised, notice that for the standard bootstrap,

$$\begin{aligned} P(\theta^* \leq a) &= \sum_{i=1}^{N} \frac{1}{N} I_{(-\infty,a]}(\theta_i) \\ &\stackrel{N\to\infty}{\longrightarrow} E_g I_{(-\infty,a]}(\theta) = \int_{-\infty}^{a} g(\theta) d\theta \end{aligned}$$

so that θ^* is approximately distributed as $g(\theta)$. For the weighted bootstrap,

$$\begin{aligned} P(\theta^* \leq a) &= \sum_{i=1}^{N} q_i I_{(-\infty,a]}(\theta_i) = \frac{\frac{1}{N}\sum_{i=1}^{N} w_i I_{(-\infty,a]}(\theta_i)}{\frac{1}{N}\sum_{i=1}^{N} w_i} \\ &\stackrel{N\to\infty}{\longrightarrow} \frac{E_g \frac{L(\theta)\pi(\theta)}{g(\theta)} \cdot I_{(-\infty,a]}(\theta)}{E_g \frac{L(\theta)\pi(\theta)}{g(\theta)}} = \frac{\int_{-\infty}^{a} L(\theta)\pi(\theta)d\theta}{\int_{-\infty}^{\infty} L(\theta)\pi(\theta)d\theta} \\ &= \int_{-\infty}^{a} h(\theta)d\theta \end{aligned}$$

so that θ^* is now approximately distributed as $h(\theta)$, as desired.

Note that, similar to the previous two indirect sampling methods, we need $g(\theta) \approx h(\theta)$, or else a very large N will be required to obtain acceptable accuracy. In particular, the "tail problem" mentioned before is potentially even more harmful here, since if there are no candidate θ_is located in the tails of $h(\theta)$ there is of course no way to resample them!

In any of the three methods discussed above, if the prior $\pi(\theta)$ is proper, it can play the role of $g(\theta)$, as shown in the following example.

Example 5.4 Suppose $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$ and $\pi(\theta) = \text{Cauchy}(\mu, \tau)$, with σ^2, μ, and τ known. The likelihood $L(\theta) \propto \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\}$ is maximized at $\hat{\theta} = \bar{y}$. Let $M = L(\hat{\theta})$ in the rejection method, and let $g(\theta) = \pi(\theta)$. Then

$$\begin{aligned} L(\theta)\pi(\theta) \leq Mg(\theta) &\iff L(\theta)\pi(\theta) \leq L(\hat{\theta})\pi(\theta) \\ &\iff L(\theta) \leq L(\hat{\theta}) . \end{aligned}$$

So we simply generate $\theta_j \sim \pi(\theta)$, $U \sim$ Uniform(0, 1), and accept θ_j if $MUg(\theta_j) < L(\theta_j)\pi(\theta_j)$, i.e., if

$$U < \frac{L(\theta_j)}{L(\hat{\theta})} .$$

Clearly this ratio is also the probability of accepting a θ_j candidate. Hence this approach will be quite inefficient unless $\pi(\theta)$ is not too flat relative to the likelihood $L(\theta)$, so that most of the candidates have a reasonable chance of being accepted. Unfortunately, this will not normally be the case, since in most applications the data will carry much more information about θ than the prior. In these instances, setting $g = \pi$ will also result in poor performance by importance sampling or the weighted bootstrap, since g will be a poor approximation to $L\pi$. As a result, g should be chosen as the prior π as a last resort, when methods of finding a better approximation have failed. ■

5.4 Markov chain Monte Carlo methods

5.4.1 Substitution sampling and data augmentation

Importance sampling, rejection sampling, and the weighted bootstrap are all "one-off," or noniterative, methods: they draw a sample of size N, and stop. Hence there is no notion of the algorithm "converging" - we simply require N sufficiently large. But for many problems, especially high-dimensional ones, it may be quite difficult or impossible to find an importance sampling density (or envelope function) which is an acceptably accurate approximation to the log posterior, but still easy to sample from.

Fortunately, some problems of this type may be fairly easy to cast in a *substitution* framework. For example, consider the three-stage hierarchical model, where, using p as a generic symbol for a density function, we have a likelihood $p(y|\theta)$, a prior $p(\theta|\eta)$, and a hyperprior $p(\eta)$. (Any or all of y, θ, and η may be vectors, but for convenience we suppress this in the notation.) The most common example of a model of this type would be the normal means model, where $y|\theta \sim N(y|\theta, 1)$, $\theta|\eta \sim N(\theta|\eta, 1)$, and $\eta \sim N(\eta|0, 1)$. Assuming the prior is conjugate with the likelihood, we can easily compute the marginal distribution,

$$p(y|\eta) = \int p(y|\theta)p(\theta|\eta)d\theta$$

so that the posterior distribution is given in closed form as

$$p(\theta|y, \eta) = \frac{p(\theta, y, \eta)}{p(y, \eta)} = \frac{p(y|\theta)p(\theta|\eta)}{p(y|\eta)} .$$

Also notice that

$$\begin{aligned} p(\eta|\theta, y) &= \frac{p(\eta, \theta, y)}{p(\theta, y)} = \frac{p(\theta|\eta)p(\eta)}{\int p(\theta|\eta)p(\eta)d\eta} \\ &= \frac{p(\theta|\eta)p(\eta)}{p(\theta)} = p(\eta|\theta) . \end{aligned}$$

Note that this distribution will also be available in closed-form given the conjugacy of the hyperprior with the prior.

For inference on θ, we seek the *marginal posterior,*

$$p(\theta|y) = \int p(\theta|y, \eta)p(\eta|y)d\eta \ ,$$

and we also have

$$p(\eta|y) = \int p(\eta|\theta)p(\theta|y)d\theta \ .$$

Switching to more generic notation, we see we have a system of two linear integral equations of the form

$$\begin{array}{rcl} p(x) & = & \int p(x|y)p(y)dy \\ p(y) & = & \int p(y|x)p(x)dx \end{array} \ , \tag{5.11}$$

where $p(x|y)$ and $p(y|x)$ are known, and we seek $p(x)$. To solve this system, we can substitute the second equation into the first and then switch the order of integration, obtaining

$$\begin{aligned} p(x) &= \int p(x|y) \int p(y|x')p(x')dx'dy \\ &= \int h(x, x')p(x')dx' \ , \end{aligned} \tag{5.12}$$

where $h(x, x') = \int p(x|y)p(y|x')dy$. Equation (5.12) is sometimes called a *fixed point* system, since if the true marginal density $p(x)$ is inserted on the right hand side, we get $p(x)$ out again on the left x' is of course just a dummy argument). But what happens if we put in *nearly* $p(x)$? That is, suppose $p_0(x)$ is some starting estimate of $p(x)$. We calculate

$$p_1(x) = \int h(x, x')p_0(x')dx' \equiv I_h(p_0(x)) \ ,$$

where I_h denotes the integral operator indexed by h. Defining $p_{i+1}(x) = I_h(p_i(x))$, $i = 0, 1, \ldots$, does this iterative process converge?

Theorem 5.2 Provided the joint density $p(x, y) > 0$ over the domain $\mathcal{X} \times \mathcal{Y}$,

(a) $p(x)$, the true marginal density, is the *unique* solution to the system (5.11).

(b) The sequence $\{p_i(x)\}$ *converges monotonically* in L_1 to $p(x)$.

(c) $\int |p_i(x) - p(x)| \, dx \to 0$ *exponentially* in i.

These results are proved by Tanner and Wong (1987) in the context of missing data problems.

Thus we have an iterative process for obtaining the true marginal density $p(x)$. But of course, the process involves performing the integration in equation (5.12) analytically, which in general will not be possible (if it were, we would never have begun this line of reasoning in the first place!). Thus instead of doing the substitution analytically, suppose we use a *sampling-based* approach, wherein we draw $X^{(0)} \sim p_0(x)$, and then draw $Y^{(1)} \sim p(y|x^{(0)})$. Note that the marginal distribution of $Y^{(1)}$ is then $p_1(y) = \int p(y|x)p_0(x)dx$. Next draw $X^{(1)} \sim p(x|y^{(1)})$, so that $X^{(1)}$ has marginal distribution

$$\begin{aligned} p_1(x) &= \int p(x|y)p_1(y)dy \\ &= \int h(x, x')p_0(x')dx' \\ &= I_h(p_0(x)) . \end{aligned}$$

By property (b) of Theorem 5.2, repeating this process produces pairs $(X^{(i)}, Y^{(i)})$ such that

$$X^{(i)} \xrightarrow{d} X \sim p(x) \quad \text{and} \quad Y^{(i)} \xrightarrow{d} Y \sim p(y) .$$

That is, for i sufficiently large, $X^{(i)}$ may be thought of as a sample from the true marginal density $p(x)$. Notice that the luxury of avoiding the integration in (5.12) has come at the price of obtaining not the marginal density $p(x)$ itself, but only a *sample* from this density. Also notice that, unlike the methods in Subsection 5.3, this algorithm is iterative, requiring sampling until "convergence" has obtained. Approaches of this type are commonly known as *Markov chain Monle Carlo* (or simply *MCMC)* methods, since their output is produced by generating samples from a certain Markov chain.

The algorithm outlined above is essentially the "data augmentation" algorithm of Tanner and Wong (1987), developed for model settings where X plays the role of the (possibly vector) parameter of interest and Y plays the role of a vector of missing data values. However, in order to obtain an estimate of the marginal density $p(x)$ (instead of merely a single sample from it), these authors complicate the algorithm slightly. At the i^{th} stage of the algorithm, they generate $Y_1^{(i)}, \ldots, Y_m^{(i)} \overset{iid}{\sim} p(y|x^{(i-1)})$, and then generate a single X value as

$$X^{(i)} \sim \hat{p}_i(x) = \frac{1}{m}\sum_{j=1}^{m} p(x|y_j^{(i)}) .$$

Here, $\hat{p}_i(x)$ is a Monte Carlo estimate of $p_i(x) = \int p(x|y)p_i(y)dy$. This produces a less variable $X^{(i)}$ sample at each iteration, but, more important it automatically produces a smooth estimate of the marginal density of interest $p(x)$ at convergence of the algorithm.

Tanner (1993) observes that $m = 1$ is sufficient for convergence of the algorithm, and refers to this approach as "chained data augmentation." In a subsequent paper, Gelfand and Smith (1990) instead recommended *parallel* sampling, as illustrated in Figure 5.4. The initial values $X_1^{(0)}, \ldots, X_m^{(0)}$ are sampled from some starting distribution $p_0(x)$, or perhaps are chosen

$$
\begin{array}{cccccc}
X_1^{(0)} & X_2^{(0)} & \cdots & X_m^{(0)} & & \\
\downarrow & \downarrow & & \downarrow & & \\
Y_1^{(1)} & Y_2^{(1)} & \cdots & Y_m^{(1)} & & \\
\downarrow & \downarrow & & \downarrow & & \\
\vdots & \vdots & & \vdots & & \\
\downarrow & \downarrow & & \downarrow & & \\
Y_1^{(i)} & Y_2^{(i)} & \cdots & Y_m^{(i)} & \rightarrow & \hat{p}_i(x) = \frac{1}{m}\sum_{j=1}^{m} p(x|y_j^{(i)})
\end{array}
$$

Figure 5.4 Diagram of parallel substitution sampling.

deterministically (say, $X_j^{(0)} = a$ for all j, where a is a naive estimate of the marginal mean of X, such as the prior mean). The advantage of this approach is that at convergence it produces m replicates that are *marginally independent,* and thus a presumably better density estimate $\hat{p}_i(x)$. An obvious disadvantage, however, is the wastefulness of the algorithm, which discards all but m of the $2im$ random variates it generates.

As an alternative, many authors (notably Geyer, 1992) have advocated the use of *ergodic* sampling. Here, we generate only one of the columns of variates illustrated in Figure 5.4, but continue sampling for an additional m - 1 iterations after convergence at iteration i. Our density estimate is then given by

$$\hat{p}_i(x) = \frac{1}{m}\sum_{j=1}^{m} p(x|y^{(i+j-1)}) \ . \tag{5.13}$$

That is, we use consecutive variates from a single long sampling chain. While these variates will not be independent, they will all come from the proper marginal distribution (the *ergodic* distribution of the chain), and hence the density estimate in (5.13) will indeed be a good approximation to $p(x)$ for i and m sufficiently large. However, the approximation will be poor if X and Y are highly correlated, since this will lead to high autocorrelation ("slow mixing") in the resulting $y^{(i)}$ sequence. That is, the chain might get "stuck" in one part of the joint distribution, leading to a density estimate based on samples that represent only a portion of the distribution. To remedy this, some authors (notably Raftery and Lewis, 1992) have recommended retaining only every k^{th} iterate after convergence has obtained, resulting in the revised ergodic density estimate

$$\hat{p}_i(x) = \frac{1}{m}\sum_{j=1}^{m} p(x|y^{(i+(j-1)k)}) \ .$$

Clearly in all of the above formulae, the proper selection of i, m, and perhaps k is critical to the success of the approach. In the interest of continuing with our presentation of the various MCMC methods, we defer further discussion of this issue to Subsection 5.4.6.

5.4.2 Gibbs sampling

While the algorithm of the previous subsection could conceivably be used in problems of dimension higher than two by thinking of X and Y as vectors, such an approach may not be feasible due to the difficulty in generating from complex multivariate distributions. As a result, we require a K-variate extension of the substitution sampling approach. For example, consider the case $K = 3$. Certainly the system of two equations in two unknowns given in (5.11) could be extended to

$$\begin{array}{rcl} p(x) & = & \int p(x, z|y)p(y)dydz \\ p(y) & = & \int p(y, x|z)p(z)dxdz \\ p(z) & = & \int p(z, y|x)p(x)dxdy \end{array} , \qquad (5.14)$$

leading to a fixed point equation of the form $p(x) = \int h(x, x'')p(x'')$, similar to (5.12). But a sampling-based version of this algorithm would require sampling from the three bivariate conditional distributions in (5.14), which are unlikely to be available in closed form. Instead, we might write the system solely in terms of univariate distributions as

$$\begin{array}{rcl} p(x) & = & \int p(x|z, y)p(z|y)p(y)dydz \\ p(y) & = & \int p(y|x, z)p(x|z)p(z)dxdz \\ p(z) & = & \int p(z|y, x)p(y|x)p(x)dxdy \end{array} . \qquad (5.15)$$

A sampling implementation could now proceed using the 6 univariate conditional distributions in (5.15), assuming that all 6 were available. Note that three of these distributions are not fully conditional, but require some marginalization beforehand - perhaps carried out by subsampling algorithms.

Now consider the general K-dimensional problem, where we have a collection of K random variables which we denote by U = $(U_1, \ldots, U_K)$. A univariate substitution approach would require sampling from a chain of $K(K - 1)$ distributions, which is likely to be impractical for large K. Fortunately, in many situations a much simpler alternative is provided by an algorithm which has come to be known as the *Gibbs sampler.* In this method, we suppose the frill or *complete* conditional distributions $\{p_i(U_i|U_{j\neq i}),\ i = 1, \ldots, K\}$ are available for sampling. (Here, "available" means that samples may be generated by some method, given values of the appropriate conditioning random variables.) Under mild conditions (see Besag, 1974), these complete conditional distributions uniquely determine the full joint distribution, $p(U_1, \ldots, U_K)$, and hence all marginal distributions $p(U_i)$, $i = 1, \ldots, K$. Then, given an arbitrary set of starting values

$\{U_1^{(0)}, \ldots, U_K^{(0)}\}$, the algorithm proceeds as follows:

$$\begin{aligned}
\text{Draw } U_1^{(1)} &\sim p_1(U_1 | U_2^{(0)}, \ldots, U_K^{(0)}), \\
\text{Draw } U_2^{(1)} &\sim p_2(U_2 | U_1^{(1)}, U_3^{(0)}, \ldots, U_K^{(0)}), \\
&\vdots \\
\text{Draw } U_K^{(1)} &\sim p_K(U_K | U_1^{(1)}, \ldots, U_{K-1}^{(1)}),
\end{aligned}$$

completing one iteration of the scheme. After t such iterations we would then obtain $(U_1^{(t)}, \ldots, U_K^{(t)})$.

Theorem 5.3 For the Gibbs sampling algorithm outlined above.

(a) $(U_1^{(t)}, \ldots U_K^{(t)}) \xrightarrow{d} (U_1 \cdots U_K) \sim p(U_1 \cdots U_K)$ as $t \to \infty$.

(b) The convergence in part (a) is *exponential* in t using the L_1 norm.

These results were originally proved for discrete distributions by Geman and Geman (1984) in the context of image restoration using Gibbs distributions (hence, the algorithm's somewhat confusing name). Schervish and Carlin (1992) provided conditions for convergence using the L_2 norm; Roberts and Smith (1993) provide conditions that are even more general and simpler to verify.

Thus, all we require to obtain samples from the joint distribution of $(U_1, \ldots U_K)$ is the ability to sample from the Ii corresponding full conditional distributions. In the context of Bayesian analysis, the joint distribution of interest is the joint *posterior* distribution, $p(U_1, \ldots, U_K|\mathbf{y})$, or perhaps one or more of the marginal posterior distributions, $p(U_i|\mathbf{y})$. The Gibbs sampler then requires draws from the full conditional posterior distributions $p(U_i|U_{j\neq i}, \mathbf{y})$. Fortunately, the conditioning inherent in these distributions implies that for a very large class of hierarchical models with conjugate priors and hyperpriors, the necessary sampling distributions will be available in closed form (see Examples 5.5 and 5.6 below). Furthermore, marginal density estimates will be available in a manner similar to that described in the previous subsection. For instance, using a few (m = 3 or 5) parallel sampling chains run for a total of T iterations each, if an acceptable degree of convergence had obtained by iteration t_0, then a marginal posterior density estimate for U_i would be given by

$$\hat{p}(U_i|\mathbf{y}) = \frac{1}{m(T-t_0)} \sum_{j=1}^{m} \sum_{t=t_0+1}^{T} p(U_i | U_{1,j}^{(t)}, \ldots, U_{i-1,j}^{(t)}, U_{i+1,j}^{(t)}, \ldots U_{K,j}^{(t)}, \mathbf{y}) \,. \tag{5.16}$$

This formula makes use of the fact that at convergence, by part (a) of Theorem 5.3 the (K-1)-tuples $(U_{1,j}^{(t)}, \ldots, U_{i-1,j}^{(t)}, U_{i+1,j}^{(t)}, \ldots U_{K,j}^{(t)})$ are distributed according to the joint posterior distribution of $(U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_K)$. We remark that the smoothing inherent in formula (5.16) leads to a less variable estimate of $p(U_i|\mathbf{y})$ than could be obtained by kernel density estimation, which would ignore the known shape of U_i's full conditional distribution. Gelfand and Smith (1990) refer to this process as "Rao-Blackwellization," since the conditioning and resulting variance reduction is reminiscent of the well-known Rao-Blackwell Theorem. It does require knowledge of the *normalized* form of the full conditional for U_i; if we lack this, we can resort to an importance-weighted marginal density estimate (see Chen, 1994) or simply a kernel density estimate (Silverman, 1986, Chapter 3).

Example 5.5 As a first, elementary illustration of the Gibbs sampler we reconsider the Poisson/gamma model of Example 2.5, as expanded in Chapter 3, Exercise 9. That is,

$$Y_i|\theta_i \overset{ind}{\sim} Poisson(\theta_i t_i),\ \theta_i \overset{ind}{\sim} G(\alpha,\beta),\ i=1,\ldots,k,$$

where the t_i are known constants. For the moment we assume that α is also known, but that β has an $IG(c, d)$ hyperprior, with known hyperparameters c and d. Thus the density functions corresponding to the three stages of the model are

$$\begin{aligned} f(y_i|\theta_i) &= \frac{e^{-(\theta_i t_i)}(\theta_i t_i)^{y_i}}{y_i!},\ y_i \geq 0,\ \theta_i > 0, \\ g(\theta_i|\beta) &= \frac{\theta_i^{\alpha-1}e^{-\theta_i/\beta}}{\Gamma(\alpha)\beta^\alpha},\ \alpha > 0,\ \beta > 0, \\ \text{and} \qquad h(\beta) &= \frac{e^{-1/(\beta d)}}{\Gamma(c)d^c\beta^{c+1}},\ c > 0,\ d > 0. \end{aligned}$$

Our main interest lies in the marginal posterior distributions of the θ_i, $p(\theta_i|\mathbf{y})$. Note that while the gamma prior is conjugate with the Poisson likelihood and the inverse gamma hyperprior is conjugate with the gamma prior, no closed form for $p(\theta_i|\mathbf{y})$ is available. However, the full conditional distributions of β and the θ_i needed to implement the Gibbs sampler are readily available. To see this, recall that by Bayes' Rule, each is proportional to the complete Bayesian model specification,

$$\left[\prod_{i=1}^{k} f(y_i|\theta_i)g(\theta_i|\beta)\right] h(\beta) \tag{5.17}$$

Thus we can find the necessary full conditional distributions by dropping irrelevant terms from (5.17) and normalizing. For θ_i we have

$$\begin{aligned} p(\theta_i|\theta_{j\neq i},\beta,\mathbf{y}) &\propto f(y_i|\theta_i)g(\theta_i|\beta) \\ &\propto \theta_i^{y_i+\alpha-1}e^{-\theta_i(t_i+1/\beta)} \\ &\propto G\left(\theta_i \mid y_i+\alpha\ ,\ (t_i+1/\beta)^{-1}\right)\ , \end{aligned} \tag{5.18}$$

while for β we have

$$\begin{aligned}
p(\beta|\{\theta_i\},\mathbf{y}) &\propto \left[\prod_{i=1}^{k} g(\theta_i|\beta)\right] h(\beta) \\
&\propto \left[\prod_{i=1}^{k} \frac{e^{-\theta_i/\beta}}{\beta^{\alpha}}\right] \frac{e^{-1/(\beta d)}}{\beta^{c+1}} \\
&\propto \frac{e^{-\frac{1}{\beta}\left(\sum_{i=1}^{k}\theta_i+\frac{1}{d}\right)}}{\beta^{k\alpha+c+1}} \\
&\propto IG\left(\beta \,\middle|\, k\alpha+c, \left(\sum_{i=1}^{k}\theta_i+1/d\right)^{-1}\right) . \qquad (5.19)
\end{aligned}$$

Thus our conditionally conjugate prior specification (stage 2 with stage 1, and stage 3 with stage 2) has allowed both of these distributions to emerge as standard forms, and the $\{\theta_i^{(t)}\}$ and $\beta^{(t)}$ may be sampled directly from (5.18) and (5.19), respectively.

As a specific illustration, we apply this approach to the pump data in Table 3.5, previously analyzed in Chapter 3, Exercise 10. These are the numbers of pump failures, Y_i, observed in t_i thousands of hours for k = 10 different systems of a certain nuclear power plant. We complicate matters somewhat by also treating α as an unknown parameter to be estimated. The complication is that its full conditional,

$$p(\alpha|\{\theta_i\},\beta,\mathbf{y}) \propto \left[\prod_{i=1}^{k} g(\theta_i|\alpha,\beta)\right] h(\alpha) ,$$

is not proportional to any standard family for any choice of hyperprior $h(\alpha)$. However, if we select an $Expo(\mu)$ hyperprior for α, then

$$p(\alpha|\{\theta_i\},\beta,\mathbf{y}) \propto \left[\prod_{i=1}^{k} \frac{\theta_i^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}\right] e^{-\alpha/\mu} ,$$

which can be shown to be log-concave in Thus we may use the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992) to update this parameter.

The **WinBUGS** code to fit this model is given in Figure 5.5. We chose the values μ = 1, c = 0.1, and d = 1.0, resulting in reasonably vague hyperpriors for α and β. **WinBUGS** recognizes the conjugate forms (5.18) and (5.19) and samples directly from them; it also recognizes the log-concavity for a and uses ARS for this parameter.

The results from running 1000 burn-in samples, followed by a "production" run of 10,000 samples (single chain) are given in Table 5.2. Note that while θ_5 and θ_6 have very similar posterior means, the latter posterior is

i	Y_i	t_i	r_i
1	5	94.320	.053
2	1	15.720	.064
3	5	62.880	.080
4	14	125.760	.111
5	3	5.240	.573
6	19	31.440	.604
7	1	1.048	.954
8	1	1.048	.954
9	4	2.096	1.910
10	22	10.480	2.099

Table 5.1 *Pump Failure Data (Gayer and O'Muircheartaigh, 1987)*

```
model
{
    for (i in 1:k) {
          theta[i] ~ dgamma(alpha,beta);
          lambda[i] <- theta[i]*t[i];
          Y[i] ~ dpois(lambda[i]);
    }
    alpha ~ dexp(1.0);
    beta ~ dgamma(0.1, 1.0);
}

#  DATA:
list(k = 10, Y = c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22),
  t = c(94.320, 15.72, 62.88, 125.76, 5.24, 31.44, 1.048,
        1.048, 2.096, 10.48))

#  INITIAL VALUES:
list(theta=c(1,1,1,1,1,1,1,1,1,1), alpha=1, beta=1)
```

Figure 5.5 *WinBUGS code for solving the pump data problem.*

much narrower (i.e., smaller posterior standard deviation). This is because, while the crude failure rates for the two pumps are similar, the latter is based on a far greater number of hours of observation $t_6 = 31.44$, while $t_5 = 5.24$). Hence we "know" more about pump 6, and this is properly reflected in its posterior distribution. ■

parameter	mean	sd	2.5%	median	97.5%
α	0.7001	0.2699	0.2851	0.6634	1.338
β	0.9290	0.5325	0.1938	0.8315	2.205
θ_1	0.0598	0.02542	0.02128	0.05627	0.1195
θ_5	0.6056	0.3150	0.1529	0.5529	1.359
θ_6	0.6105	0.1393	0.3668	0.5996	0.9096
θ_{10}	1.993	0.4251	1.264	1.958	2.916

Table 5.2 *Posterior summaries, pump data problem*

Example 5.6 Our second Gibbs sampling example concerns a longitudinal dataset on the weights of young laboratory rats, given in Table 5.3 and originally considered by Gelfand, Hills, Racine-Poop, and Smith (1990). While this dataset is already something of a "golden oldie" in the MCMC Bayesian literature, we include it since it is a problem of moderate difficulty that exploits the sampler in the context of the normal linear model, common in practice. In these data, Y_{ij} corresponds to the weight of the i^{th} rat at measurement point *j*, while x_{ij} denotes its age in days at this point. We adopt the model

$$Y_{ij} \overset{ind}{\sim} N\left(\alpha_i + \beta_i x_{ij}\ ,\ \sigma^2\right),\ i = 1, \ldots, k,\ j = 1, \ldots, n_i\ , \tag{5.20}$$

where $k = 30$ and $n_i = 5$ for all i (unlike humans, laboratory rats can be counted upon to show up for *all* of their office visits!). Thus the model supposes a straight-line growth curve with homogeneous errors, but allows for the possibility of a *different* slope and intercept for each individual rat. Still, we believe all the growth curves to be similar, which we express statistically by assuming that the slope-intercept vectors are drawn from a common normal population, i.e.,

$$\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}\ ,\ \Sigma\right), \quad i = 1, \ldots, k\ . \tag{5.21}$$

Thus we have what a classical statistician might refer to as a *random effects model* (or a *random coefficient model*).

To effect a Bayesian analysis, we must complete the hierarchical structure at the second stage with a prior distribution for σ^2 and at the third stage with prior distributions for $\boldsymbol{\theta}_0$ and Σ. We do this by choosing conjugate forms for each of these distributions, namely,

$$\begin{aligned} \sigma^2 &\sim IG(a, b)\ , \\ \boldsymbol{\theta}_0 &\sim N(\boldsymbol{\eta}, C)\ , \text{ and} \\ \Sigma^{-1} &\sim W\left((\rho R)^{-1}, \rho\right)\ , \end{aligned} \tag{5.22}$$

Rat i	Weight measurements Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}
1	151	199	246	283	320
2	145	199	249	293	354
3	147	214	263	312	328
4	155	200	237	272	297
5	135	188	230	280	323
6	159	210	252	298	331
7	141	189	231	275	305
8	159	201	248	297	338
9	177	236	285	340	376
10	134	182	220	260	296
11	160	208	261	313	352
12	143	188	220	273	314
13	154	200	244	289	325
14	171	221	270	326	358
15	163	216	242	281	312
16	160	207	248	288	324
17	142	187	234	280	316
18	156	203	243	283	317
19	157	212	259	307	336
20	152	203	246	286	321
21	154	205	253	298	334
22	139	190	225	267	302
23	146	191	229	272	302
24	157	211	250	285	323
25	132	185	237	286	331
26	160	207	257	303	345
27	169	216	261	295	333
28	157	205	248	289	316
29	137	180	219	258	291
30	153	200	244	286	324

Table 5.3 *Rat population growth data (Gelfand et al., 1990, Table 3).*

where W denotes the Wishart distribution, a multivariate generalization of the gamma distribution. Here, R is a 2 x 2 matrix and $\rho \geq 2$ is a scalar "degrees of freedom" parameter. In more generic notation, a $p \times p$ symmetric and positive definite matrix V distributed as $W(D, n)$ has density function proportional to

$$\frac{|V|^{(n-p-1)/2}}{|D|^{n/2}} \exp\left[-\frac{1}{2} tr(D^{-1}V)\right] \tag{5.23}$$

provided that $n \geq p$. For this distribution, one can show $E(V_{ij}) = nD_{ij}$, $Var(V_{ij}) = n(D_{ij}^2 + D_{ii}D_{jj})$, and $Cov(V_{ij}, V_{kl}) = n(D_{ik}D_{jl} + D_{il}D_{jk})$. Thus under the parametrization chosen in (5.22) we have $E(\Sigma^{-1}) = R^{-1}$, so R^{-1} is the expected prior precision of the $\boldsymbol{\theta}_i$s (i.e., R is approximately the expected prior variance). Also, $Var(\Sigma_{ij})$ is decreasing in ρ, so small p values correspond to vaguer prior distributions. For more on the Wishart distribution, the reader is referred to Press (1982, Chapter 5).

The hyperparameters in our model are a, b, $\boldsymbol{\eta}$, C, p, and R, all of which are assumed known. We seek the marginal posterior for $\boldsymbol{\theta}_0$ given only the observed data. The total number of unknown parameters in our model is 66: 30 α_is, 30 β_is, α_0, β_0, σ^2, and the 3 unique components of Σ. This number is far too high for any approximation or numerical quadrature method, and the noniterative Monte Carlo methods in Section 5.3 would be very cumbersome. By contrast, the Gibbs sampler is relatively straightforward to implement here thanks to the conjugacy of the prior distributions at each stage in the hierarchy. For example, in finding the full conditional for $\boldsymbol{\theta}_i$ we are allowed to think of the remaining parameters as fixed. But as proved by Lindley and Smith (1972) and shown in Example 2.7, with variance parameters known the posterior distribution for $\boldsymbol{\theta}_i$ emerges as

$$\boldsymbol{\theta}_i | \mathbf{y}, \boldsymbol{\theta}_0, \Sigma^{-1}, \sigma^2 \sim N\left(D_i(\sigma^{-2} X_i^T \mathbf{y}_i + \Sigma^{-1}\boldsymbol{\theta}_0) \, , \, D_i\right)$$

for $i = 1, \ldots, k$, where $D_i^{-1} = \sigma^{-2} X_i^T X_i + \Sigma^{-1}$,

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad \text{and} \quad X_i = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{pmatrix} .$$

Similarly, the full conditional for $\boldsymbol{\theta}_0$ is

$$\boldsymbol{\theta}_0 | \mathbf{y}, \{\boldsymbol{\theta}_i\}, \Sigma^{-1}, \sigma^2 \sim N\left(V(k\Sigma^{-1}\bar{\boldsymbol{\theta}} + C^{-1}\boldsymbol{\eta}) \, , \, V\right) ,$$

where V = $(k\Sigma^{-1} + C^{-1})^{-1}$ and $\bar{\boldsymbol{\theta}} = \frac{1}{k}\sum_{i=1}^k \boldsymbol{\theta}_i$.

For Σ^{-1}, under the parametrization in (5.22), the prior distribution (5.23) is proportional to

$$|\Sigma^{-1}|^{(\rho-3)/2} \exp\left[-\frac{1}{2} tr(\rho R \Sigma^{-1})\right] .$$

Combining the normal likelihood for the random effects (5.21) with this prior distribution produces the updated Wishart distribution

$$\begin{aligned} &\Sigma^{-1} | \mathbf{y}, \{\boldsymbol{\theta}_i\}, \boldsymbol{\theta}_0, \sigma^2 \\ &\qquad \sim W\left(\left[\textstyle\sum_{i=1}^k (\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)^T + \rho R\right]^{-1}, \, k + \rho\right) . \end{aligned}$$

Finally, the full conditional for σ^2 is the updated inverse gamma distribution

$$\sigma^2 \big| \mathbf{y}, \{\boldsymbol{\theta}_i\}, \boldsymbol{\theta}_0, \Sigma \\ \sim IG\left(\frac{n}{2} + a\,,\ \left[\frac{1}{2}\sum_{i=1}^k (\mathbf{y}_i - X_i\boldsymbol{\theta}_i)^T(\mathbf{y}_i - X_i\boldsymbol{\theta}_i) + b^{-1}\right]^{-1}\right)\,,$$

where n = $\sum_{i=1}^k n_i$. Thus the conditioning property of the Gibbs sampler enables closed form full conditional distributions for each parameter in the model, including the variance parameters - the sticking point in the original Lindley and Smith paper. Samples from all of these conditional distributions (except perhaps the Wishart) are easily obtained from any of a variety of statistical software packages. Wishart random deviates may be produced via an appropriate combination of normal and gamma random deviates. The reader is referred to the original paper on this topic by Odell and Feiveson (1966), or to the more specifically Bayesian treatment by Gelfand et al. (1990).

In our dataset, each rat was weighed once a week for five consecutive weeks. In fact, the rats were all the same age at each weighing: $x_{i1} = 8$, $x_{i2} = 15$, $x_{i3} = 22$, $x_{i4} = 29$, and $x_{i5} = 36$ for all i. As a result, we may simplify our computations by rewriting the likelihood (5.20) as

$$Y_{ij} \stackrel{ind}{\sim} N\left(\alpha_i + \beta_i(x_{ij} - \bar{x})\,,\ \sigma^2\right)\,,\ i = 1, \ldots, k,\ j = 1, \ldots, n_i\,,$$

so that it is now reasonable to think of α_i and β_i as independent a priori. Thus we may set $\Sigma = Diag(\sigma_\alpha^2, \sigma_\beta^2)$, and replace the Wishart prior (5.22) with a product of independent inverse gamma priors, say $IG(a_\alpha, b_\alpha)$ and $IG(a_\beta, b_\beta)$.

We complete the prior specification by choosing hyperprior values that determine very vague priors, namely, C^{-1} = 0 (so that $\boldsymbol{\eta}$ disappears entirely from the full conditionals), $a = a_\alpha = a_\beta = \epsilon$, and $b = b_\alpha = b_\beta = 1/\epsilon$ where ϵ = 0.001. Using the **BUGS** language (Spiegelhalter et al., 1995a), we ran three independent Gibbs sampling chains for 500 iterations each. These chains were started from three different points in the sample space that we believed to be overdispersed with respect to the posterior distribution, so that overlap among these chains would suggest convergence of the sampler. (We hasten to remind the reader that diagnosing convergence of MCMC algorithms is a very difficult problem, and is the subject of much further discussion in Subsection 5.4.6 below.)

We summarized the Gibbs samples generated by **BUGS** using **CODA** (Best, Cowles, and Vines, 1995), a menu-driven collection of *S*+ routines useful for convergence diagnosis and output analysis. (See the Appendix Subsection C.3.3 MCMC software guide for more information on the acquisition and use of **BUGS** and **CODA**.) The first column of Figure 5.6 displays the Gibbs samples obtained for the population slope and intercept at time zero (birth), the latter now being given by $\mu = \alpha_0 - \beta_0\bar{x}$ due to our recentering of the x_{ij}s. As a satisfactory degree of convergence seems to have occurred by

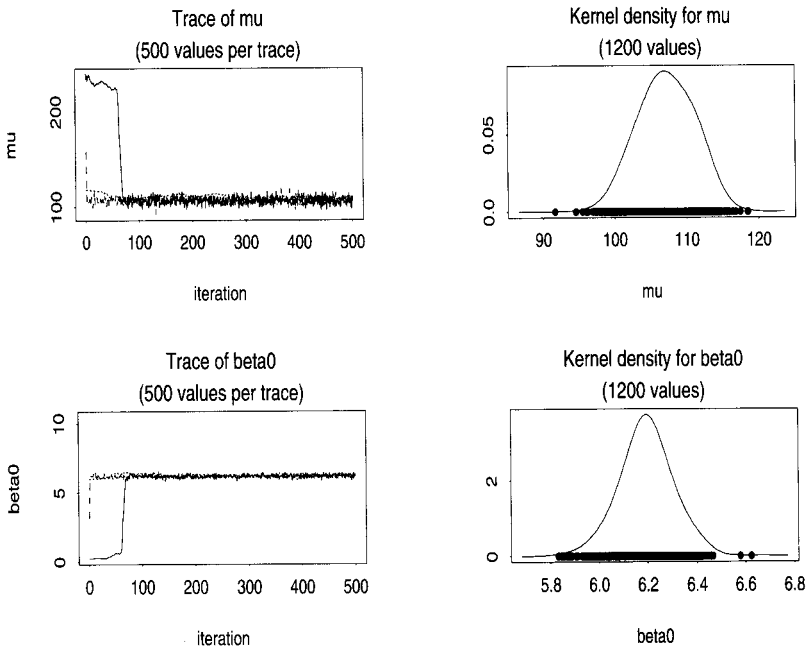


Figure 5.6 *Monitoring plots and kernel density estimates for* μ and β_0*, Gibbs sampling analysis of the rat data.*

iteration 100, we use the output of all three chains over iterations 101-500 to obtain the posterior density estimates shown in the second column of the figure. These are only kernel density estimates; more accurate results could be obtained via the Rao-Blackwellization approach in equation (5.16). It appears that the average rat weighs about 106 grams at birth, and gains about 6.2 grams per day.

Gelfand et al. (1990) also discuss and illustrate the impact of missing data on the analysis. If Y_{ij} is missing, samples from its predictive distribution $p(y_{ij}|\mathbf{y})$ are easily generated, since its complete conditional distribution is nothing but the normal distribution given in the likelihood expression (5.20). The predictive could then be estimated using a kernel smooth of these samples as in Figure 5.6, or a mixture density estimate of the form (5.16), namely,

$$
\begin{aligned}
p(y_{ij}|\mathbf{y}) &= \int p(y_{ij}|\boldsymbol{\theta}_i, \sigma^2) p(\boldsymbol{\theta}_i, \sigma^2|\mathbf{y}) \\
&\approx \frac{1}{G}\sum_{g=1}^{G} p\left(y_{ij}|\boldsymbol{\theta}_i^{(g)}, \sigma^{2(g)}\right) ,
\end{aligned}
$$

where the superscript (g) indexes the post-convergence replications from the sampler. ■

In the previous example, the prior distribution of each parameter was chosen to be conjugate with the corresponding likelihood term that preceded it in the hierarchy. This enabled all of the full conditional distributions necessary for implementing the Gibbs sampler to emerge in closed form as members of familiar distributional families, so that sample generation was straightforward. But in many examples, we may want to select a prior that is not conjugate; in others (e.g., nonlinear growth curve models), no conjugate prior will exist. How are we to perform the required sampling in such cases?

Provided that the model specification (likelihood and priors) are available in closed form, any given full conditional will be available at least up to a constant of proportionality, since

$$p(\theta_i|\{\theta_{j\neq i}\}, \mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ . \tag{5.24}$$

Hence the "nonconjugacy problem" for Gibbs sampling outlined in the previous paragraph is nothing but a univariate version of the basic computation problem of sampling from nonstandardized densities. This suggests using one of the indirect methods given in Subsection 5.3.2 to sample from these full conditionals. Unfortunately, the fact that this problem is embedded within a Gibbs sampler makes some of these techniques prohibitively expensive to implement. For example, recall that the weighted bootstrap operates by resampling from a finite sample from an approximation to the target distribution with appropriately determined unequal probabilities of selection. Sampling from (5.24) at the t^{th} iteration of the Gibbs algorithm would thus require drawing a large approximating sample, of which only one would be retained. Worse yet, this initial sample could not typically be reused, since the values of the conditioning random variables $\{\theta_{j\neq i}\}$ will have changed by iteration (t + 1).

Rejection sampling will have similarly high start-up costs when applied in this context, though several ingenious modifications to reduce them have been proposed in the MCMC Bayesian literature. Wakefield et al. (1991) developed a generalized version of the ratio-of-uniforms method (see e.g. Devroye, 1986, p.194), a tailored rejection method that is particularly useful for sampling from nonstandardized densities with heavy tails. Gilks and Wild (1992) instead use traditional rejection, but restrict attention to log-concave densities (i.e., ones for which the second derivative of the log of (5.24) with respect to θ_i is nonpositive everywhere). For these densities, an envelope on the log scale can be formed by two lines, one of which is tangent to log p left of the mode, and the other tangent right of the mode. The exponential of this piecewise linear function is thus guaranteed to envelope exp(log p) = p. In a subsequent paper, Gilks (1992) developed a derivative-free version of this algorithm based on secant (instead of tangent) lines.

The log-concavity property can be a nuisance to check, but is typically not a serious restriction (though the *t* family is a notable exception).

In an attempt to avoid having to "re-envelope" the target density at each iteration, Carlin and Gelfand (1991b) recommended a multivariate rejection algorithm. In this approach, a single multivariate split-normal or split-*t* envelope function is obtained for $L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ at the algorithm's outset. Cross-sectional slices of this envelope in the appropriate coordinate direction may then serve as univariate envelopes for the full conditional distributions regardless of the values of the conditioning random variables at a particular iteration. Unfortunately, while this approach greatly reduces the amount of computational overhead in the algorithm, the difficulty in finding an envelope whose shape even vaguely resembles that of $L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ across $\boldsymbol{\theta}$ limits its applicability to fairly regular surfaces of moderate dimension.

Yet another alternative is provided by Ritter and Tanner (1992), who, instead of rejection, recommend approximate inversion using a grid-based estimate of full conditional cdf, $\hat{F}(\theta_i|\{\theta_{j\neq i}\}$ y). Since $F^{-1}(U)$ will have distribution F when U is distributed Uniform(0, 1), $\hat{F}^{-1}(U)$ should have distribution Wroximately equal to F, with the quality of the approximation improving as the mesh of the grid goes to zero. Unfortunately, the establishment of $\hat{F}$ is very numerically intensive, involving a large number of function evaluations, and the quality of the algorithm's output is difficult to assess.

In summary, all of these solutions seem rather tedious to program, involve a large amount of computational overhead, and are of questionable accuracy (the rejection methods due to possible envelope violations; the inversion method due to the piecewise linear nature of the approximate cdf). Perhaps when we encounter a substantial number of unstandardized full conditionals, the problem is trying to tell us something, and we should look for an alternate Markov chain Monte Carlo method. The next subsection investigates such a method. Subsection 5.4.4 goes on to show how the two approaches may be combined to form what is possibly the best (and certainly the easiest) solution to the nonconjugacy problem.

5.4.3 Metropolis-Hastings algorithm

Like the Gibbs sampler, the Metropolis-Hastings algorithm is an MCMC method that has been widely used in the mathematical physics and image restoration communities for several years, but only recently discovered by statisticians. The method in its simplest form dates to the paper by Metropolis et al. (1953), which we now summarize briefly. Suppose the true joint posterior for a (possibly vector-valued) parameter U has density $p(u)$ with respect to some dominating measure μ. Choose an auxiliary function $q(v, u)$ such that $q(., u)$ is a pdf with respect to μ for all u, and $q(u, v) = q(v, u)$ for all u, v (that is, q is symmetric in its arguments).

The function q *is* often called the *candidate* or *proposal* density. We then generate a Markov chain as follows:

1. Draw v ~ q(-, u), where $u = U^{(t-1)}$, the current state of the Markov chain;
2. Compute the ratio r = p(v)/p(u);
3. If r ≥ 1, set $U^{(t)} = v$;

   If r < 1, set $U^{(t)} = \begin{cases} v \text{ with probability } r \\ u \text{ with probability } 1-r \end{cases}$.

Notice that the joint posterior density p is needed only up to proportionality constant, since it is only used to compute the acceptance ratio in step 2. But, as already noted, in Bayesian applications $p(u) \propto L(u)\pi(u)$, a form that is typically available.

Theorem 5.4 For the Metropolis algorithm outlined above, under mild conditions, $U^{(t)} \xrightarrow{d} U \sim p$ as $t \to \infty$.

The general proof of this result using the theory of irreducible Markov chains (Nummelin, 1984) is given by Tierney (1994). For finite state spaces, Ripley (1987, p.113) shows why we need symmetry in q. In this case, the transition kernel can be represented by a matrix, say P, wherein P_{ij} gives the probability of moving from state i to state j. Our Markov chain will have equilibrium distribution $\mathbf{d} = (d_1, \ldots, d_k)'$ if and only if

$$\mathbf{d}'P = \mathbf{d}' \; .$$

Since our candidate transition matrix Q is symmetric, we have

$$\begin{aligned} d_i P_{ij} &= d_i \left[\min\left(1, \frac{d_j}{d_i}\right) Q_{ij} \right] \\ &= \min(d_i, d_j) Q_{ij} \\ &= \min(d_i, d_j) Q_{ji} \\ &= d_j P_{ji} \; . \end{aligned}$$

That is, the chain is *reversible.* But this in turn implies that

$$\begin{aligned} (\mathbf{d}'P)_j &= \sum_{i=1}^{k} d_i P_{ij} \;=\; \sum_{i=1}^{k} d_j P_{ji} \\ &= d_j \sum_{i=1}^{k} P_{ji} \;=\; d_j \; . \end{aligned}$$

Thus our Metropolis chain does have the proper equilibrium distribution **d**.

To show convergence to and uniqueness of the equilibrium distribution, we need a bit more machinery. The equilibrium distribution is unique if P is *irreducible*; that is, if for every set A with μ(A) > 0, the probability of the chain ever entering A is positive for every starting point u. (Note that P is irreducible if and only if Q is.) The distribution of $U^{(t)}$ will converge to the equilibrium distribution if and only if P is *aperiodic*. That is, the chain can move from any state to any other; there can be no "absorbing" states (or collection of states) from which escape is impossible.

For the continuous parameter settings we shall encounter, the most convenient choice for q is a $N(\boldsymbol{\theta}^{(t-1)}, \tilde{\Sigma})$ density, since it is easily sampled and clearly symmetric in $\mathbf{v}$ and $\boldsymbol{\theta}^{(t-1)}$. While in theory any positive definite covariance matrix $\tilde{\Sigma}$ should suffice, in practice its selection requires a bit of care. Thinking in one dimension for the moment notice that choosing too large a value for $\tilde{\Sigma}$ will cause the candidate density to take large jumps around the parameter space - many to places far from the support of the posterior. Thus many of these candidates will be rejected, and the chain will have a tendency to "get stuck" in one state and not traverse the sample space well. On the other hand, choosing an overly small $\tilde{\Sigma}$ can be harmful as well, since then the chain will have a high acceptance rate, but wind up "baby-stepping" around the parameter space. This will again increase the autocorrelation in the chain and lengthen the time required to visit the entire space.

Müller (1991) suggested matching $\tilde{\Sigma}$ as nearly as possible to the true posterior covariance matrix, perhaps by obtaining a crude estimate of $\tilde{\Sigma}$ from an initial sampling run. Other authors have recommended choosing $\tilde{\Sigma}$ to provide an observed acceptance ratio near 50%. A recent investigation by Gelman, Roberts, and Gilks (1996) shows that this second piece of advice is close to optimal when the target density is univariate normal: the first order autocorrelation in the chain is minimized by taking $\sqrt{\tilde{\Sigma}}$ equal to 2.4 times the true posterior standard deviation, and this candidate density leads to a 44.1% acceptance rate. However, matters are a bit different in higher dimensions, as these authors go on to show using the theory of Langevin diffusion processes. For the multivariate target density having product form

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{K} g(\theta_i)$$

for some one-dimensional density g, the optimal acceptance rate approaches 23.4% as the dimensionality (K) goes to infinity. This suggests that in high dimensions, it is worth risking an occasional sticking point in the algorithm in order to gain the benefit of an occasional large jump across the parameter space.

A simple but important generalization of the Metropolis algorithm was provided by Hastings (1970). This paper dropped the requirement that q

be symmetric, and redefined the acceptance ratio as

$$r = \frac{p(v)q(u,\, v)}{p(u)q(v,\, u)} \,. \tag{5.25}$$

It is easy to show that with this small modification, the revised algorithm now converges to the desired target distribution for any candidate density q. For example, we might let $V = u + Z$ where Z is distributed according to some density g. That is, we have $q(v, u) = g(v - u)$; if g is symmetric about the origin, this reduces to the original Metropolis algorithm. Tierney (1994) refers to a chain produced via this approach as a *random walk chain.* Alternatively, we might employ an *independence chain* wherein $V = Z$, so that $q(v, u) = g(v)$ (i.e., we completely ignore the current value of the chain when generating the candidate). Here, using equation (5.25) we obtain the acceptance ratios $r = w(v)/w(u)$ where $w(u) = p(u)/g(u)$, the usual weight function used in importance sampling. This suggests using the same guidelines for choosing g as for choosing a good importance sampling density (i.e., making g a good match for p), but perhaps with heavier tails to obtain an acceptance rate closer to 0.5, as encouraged by Gelman, Roberts and Gilks (1996). Further guidance on the proper choice of Hastings kernel, as well as a derivation of the algorithm from the logic of reversibility, is given by Chib and Greenberg (1995a).

Example 5.7 We illustrate the Metropolis-Hastings algorithm using the data in Table 5.4, which are taken from Bliss (1935). These data record the number of adult flour beetles killed after five hours of exposure to various levels of gaseous carbon disulphide (CS2). Since the variability in these data cannot be explained by the standard logistic regression model, we attempt to fit the generalized logit model suggested by Prentice (1976),

$$P(\text{death}|w) \equiv h(w) = \{\exp(x)/(1+\exp(x))\}^{m_1} \,.$$

Here, w is the predictor variable (dose), and $x = (w - \mu)/\sigma$ where $\mu \in \Re$ and σ^2, $m_1 > 0$. Suppose there are y_i flour beetles dying out of n_i exposed at level w_i i = I-- . . , k.

For our prior distributions, we assume that $m_1 \sim \text{Gamma}(a_0, b_0)$, $\mu \sim \text{Normal}(c_0, d_0)$, and $\sigma^2 \sim$ Inverse Gamma(e_0, f_0), where a_0, b_0, c_0, d_0, e_0, and f_0 are known, and m_1, μ, and σ^2 are independent. While these common families may appear to have been chosen to preserve some sort of conjugate structure, this is not the case: there is no closed form available for any of the three full conditional distributions needed to implement the Gibbs sampler. Thus we instead resort to the Metropolis algorithm. Our likelihood-prior specification implies the joint posterior distribution

$$\begin{aligned} p(\mu, \sigma^2, m_1|\mathbf{y}) &\propto f(\mathbf{y}|\mu, \sigma^2, m_1)\pi(\mu, \sigma^2, m_1) \\ &\propto \left\{ \prod_{i=1}^{k} [h(w_i)]^{y_i}[1 - h(w_i)]^{n_i - y_i} \right\} \frac{m_1^{a_0 - 1}}{\sigma^{2(e_0+1)}} \end{aligned}$$

Dosage w_i	# killed y_i	# exposed n_i
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60

Table 5.4 *Flour beetle mortality data*

$$\times \exp\left[-\frac{1}{2}\left(\frac{\mu - c_0}{d_0}\right)^2 - \frac{m_1}{b_0} - \frac{1}{f_0\sigma^2}\right].$$

We begin by making a change of variables from (μ, σ^2, m_1) to $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\mu, \frac{1}{2}\log\sigma^2, \log m_1)$. This transforms the parameter space to $\Re^3$ (necessary if we wish to work with Gaussian proposal densities), and also helps to symmetrize the posterior distribution. Accounting for the Jacobian of this transformation, our target density is now

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad & \left\{\prod_{i=1}^{k}[h(w_i)]^{y_i}[1-h(w_i)]^{n_i-y_i}\right\}\exp(a_0\theta_3 - 2e_0\theta_2) \\ & \times \exp\left[-\frac{1}{2}\left(\frac{\theta_1 - c_0}{d_0}\right)^2 - \frac{\exp(\theta_3)}{b_0} - \frac{\exp(-2\theta_2)}{f_0}\right]. \end{aligned}$$

Numerical stability is improved by working on the log scale, i.e., by computing the Metropolis acceptance ratio as $r = \exp[\log p(\mathbf{v}|\mathbf{y}) - \log p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})]$.

We complete the prior specification by choosing the same hyperparameter values as in Carlin and Gelfand (1991b). That is, we take a_0 = .25 and b_0 = 4, so that m_1 has prior mean 1 (corresponding to the standard logit model) and prior standard deviation 2. We then specify rather vague priors for μ and σ^2 by setting c_0 = 2, d_0 = 10, e_0 = 2.000004, and f_0 = 1000; the latter two choices imply a prior mean of .001 and a prior standard deviation of .5 for σ^2. Figure 5.7 shows the output from three parallel Metropolis sampling chains, each run for 10,000 iterations using a $N_3(\boldsymbol{\theta}^{(t-1)}, \widetilde{\Sigma})$ proposal density where

$$\widetilde{\Sigma} = D = Diag(.00012, .033, .10)\,. \tag{5.26}$$

The histograms in the figure include all 24,000 samples obtained after a burn-in period of 2000 iterations (reasonable given the appearance of the

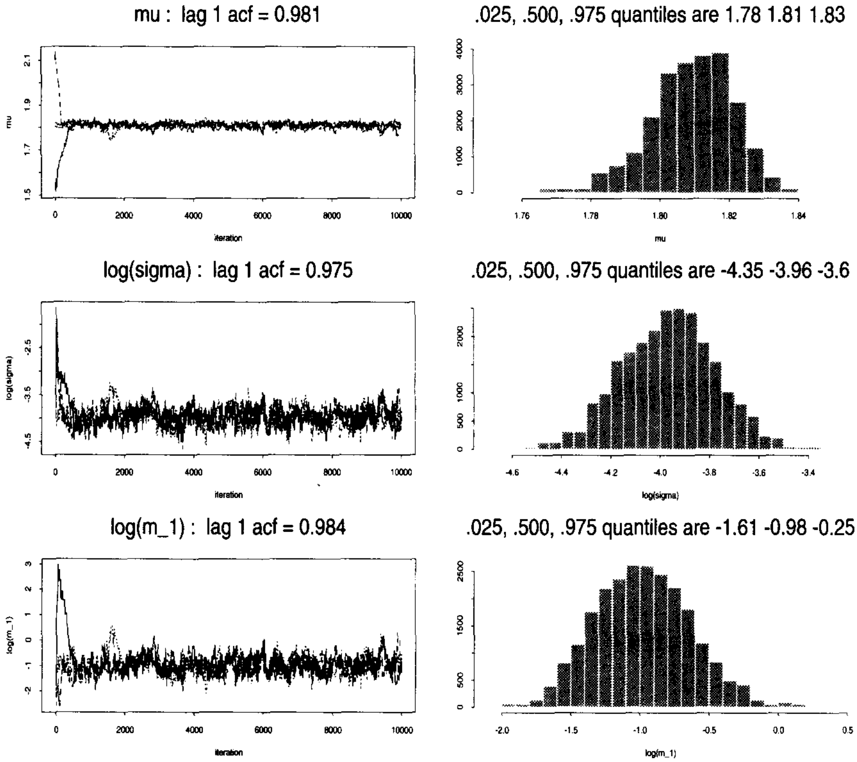


Figure 5.7 *Metropolis analysis of the flour beetle mortality data using a Gaussian proposal density with a diagonal* $\widetilde{\Sigma}$ *matrix. Monitoring plots use three parallel chains, and histograms use all samples following iteration 2000. Overall Metropolis acceptance rate: 13.5%.*

monitoring plots). The chains mix very slowly, as can be seen from the extremely high lag 1 sample autocorrelations, estimated from the chain #2 output and printed above the monitoring plots. The reason for this slow convergence is the high correlations among the three parameters, estimated as $\widehat{Corr}(\theta_1, \theta_2) = -0.78$, $\widehat{Corr}(\theta_1, \theta_3) = -0.94$, and $\widehat{Corr}(\theta_2, \theta_3) = 0.89$ from the chain #2 output. As a result, the proposal acceptance rate is low (13.5%) and convergence is slow, despite the considerable experimentation in arriving at the three variances in our proposal density.

We can accelerate convergence by using a nondiagonal proposal covariance matrix designed to better mimic the posterior surface itself. From the output of the first algorithm, we can obtain an estimate of the posterior covariance matrix in the usual way as $\widehat{\Sigma} = \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}})'$, where j indexes Monte Carlo samples. (We used the IMSL routine DCORVC for

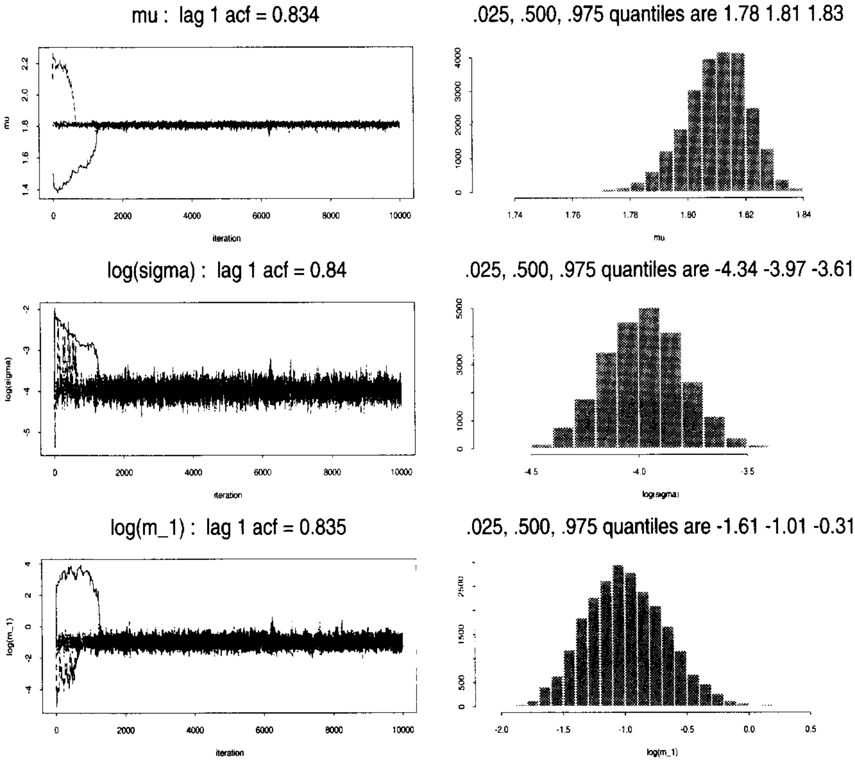


Figure 5.8 *Metropolis analysis of the flour beetle mortality data using a Gaussian proposal density with a nondiagonal* $\tilde{\Sigma}$ *matrix. Monitoring plots use three parallel chains, and histograms use all samples following iteration 2000. Overall Metropolis acceptance rate: 27.3%.*

this purpose.) We then reran the algorithm with proposal variance matrix

$$\tilde{\Sigma} = 2\hat{\Sigma} = \begin{pmatrix} 0.000292 & -0.003546 & -0.007856 \\ -0.003546 & 0.074733 & 0.117809 \\ -0.007856 & 0.117809 & 0.241551 \end{pmatrix} . \tag{5.27}$$

The results, shown in Figure 5.8, indicate improved convergence, with lower observed autocorrelations and a higher Metropolis acceptance rate (27.3%). This rate is close to the optimal rate derived by Gelman, Roberts, and Gilks (1996) of 31.6%, though this rate applies to target densities having three *independent* components. The marginal posterior results in both figures are consistent with the results obtained for this prior-likelihood combination by Carlin and Gelfand (1991b) and Miller (1991). ■

5.4.4 *Hybrid forms and other algorithms*

One of the nice features of MCMC algorithms is that they may be combined in a single problem, in order to take advantage of the strengths of each. To illustrate, suppose that $P_1, \ldots, P_m$ are Markov kernels all having stationary distribution p. These kernels might correspond to a Gibbs sampler and (m-1) Metropolis-Hastings algorithms, each with a different proposal density. A *mixture* of the algorithms arises when at iteration t, kernel P_i is chosen with probability α_i, where $\sum_{i=1}^{m} \alpha_i = 1$. Alternatively, we might simply use the kernels in a *cycle*, wherein each kernel P_i is used in a prespecified order. Clearly, a mixture or cycle of convergent algorithms will also produce a convergent algorithm.

Example 5.8 Let P_1 correspond to a standard Gibbs sampler, and P_2 correspond to an independence chain Hastings algorithm (i.e., one where $q(v, u) = g(v)$, independent of u). Suppose g is overdispersed with respect to the true posterior. For example, g might have roughly the same mode as p, but twice as large a standard deviation. Then a mixture of P_1 and P_2 with (α_1, α_2) = (.9, .1) will serve to occasionally "jump start" the sampler, possibly with a large leap across the parameter space. This will in turn reduce autocorrelations while still preserving the convergence to the stationary distribution p. Perhaps more simply, P_2 could be used once every 10 or 100 iterations in a deterministic cycle. ■

Another profitable method of combining transition kernels is motivated by the nonconjugacy problem mentioned at the end of Subsection 5.4.2. Suppose that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, and that all of the full conditional distributions are available in closed form *except* for $p_i(\theta_i | \theta_{j \neq i}, \mathbf{y})$, perhaps due to nonconjugacy in the prior for θ_i. Since the Metropolis-Hastings algorithm does not require normalized densities from which to sample, we might naturally think of running a Gibbs sampler with a Metropolis *subalgorithm* to obtain the necessary $\theta_i^{(t)}$ sample at iteration *t*. That is, we would choose a symmetric *conditional* candidate density

$$q\left(v, \theta_i^{(t-1)} \;\middle|\; \theta_{j<i}^{(t)},\, \theta_{j>i}^{(t-1)}, \mathbf{y}\right) ,$$

and run a Metropolis subalgorithm for T iterations, accepting or rejecting the candidates as appropriate. We would then set $\theta_i^{(t)}$ equal to the end result of this subalgorithm, and proceed on with the outer Gibbs loop. Note that this same approach could be applied to any parameter for which we lack a closed form full conditional. This approach is often referred to as *Metropolis within Gibbs,* though some authors (notably Besag et al., 1995, and Chib and Greenberg, 1995a) dislike this term, noting that Metropolis et al. (1953) and Hastings (1970) envisioned such algorithms in their original papers - long before the appearance of the Gibbs sampler. As a result, the terms *univariate Metropolis* or simply *Metropolis steps* might be used, to distinguish the approach from Gibbs steps (where the candidate is always accepted) and the multivariate Metropolis algorithm introduced in Subsection 5.4.3.

In using such an algorithm, two convergence issues arise. The first is a theoretical one: does this hybrid algorithm actually converge to the correct stationary distribution? The algorithm does not quite meet our definition of a cycle, since it is a deterministic. combination of Gibbs and Metropolis steps that would *not* themselves converge if applied individually (since the former does not update θ_i, while the latter does not update the others). But each of the components *does* preserve the stationary distribution of the chain, so, provided the hybrid chain is aperiodic and irreducible, convergence still obtains.

The second issue is more practical: how should we select T, the number of Metropolis subiterations at each stage of the outer algorithm? We have so far been vague about how to terminate the Gibbs sampler, saying only that it is a hard problem; does not this hybrid algorithm then embed a hard problem within a hard problem? Fortunately, the answer is no: diagnosing convergence is no harder (or easier) than before. This is because convergence will occur for *any* value of T, and so we are free to pick whatever value we like. This is not to say that in a particular problem, any T value will be as good as any other with respect to convergence. T = 100, 000 would no doubt be a poor choice, since all we would gain from this extensive sampling is high confidence that $\theta_i^{(t)}$ is indeed a sample from the correct full conditional distribution. This would be of little value early on (i.e., when t is small), since the outer Gibbs algorithm would still be far from locating the true stationary distribution, and so the expensive sample would not likely be retained for subsequent inference. On the other hand, T = 1 might correspond to such a crude approximation to the true full conditional that convergence of the overall algorithm is retarded. In a trivariate normal setting investigated by Gelfand and Lee (1993), T = 4 or 8 provided good performance, but of course this does not constitute a general recommendation. In practice, T = 1 is often adopted, effectively substituting one Metropolis rejection step for each Gibbs step lacking an easily sampled complete conditional.

Müller (1991) suggested a specific implementation of this idea intended to speed convergence. Besides employing univariate Metropolis substeps for generation from unstandardized complete conditionals, it also employs a simple transformation that attempts to orthogonalize the parameter space. Specifically, at iteration t let $\widehat{\Sigma} = K K'$ be a current estimate and Cholesky decomposition of the covariance matrix of $\boldsymbol{\theta}$. Let $\boldsymbol{\eta}^{(t)} = K^{-1}\boldsymbol{\theta}^{(t)}$ and run a univariate Metropolis algorithm on the components of $\boldsymbol{\eta}$ instead of $\boldsymbol{\theta}$. if $\boldsymbol{\theta}|\mathbf{y})$ resembles a normal distribution, then the components of $\boldsymbol{\eta}$ should be approximately uncorrelated, thus speeding convergence of the sampler.

Miner recommends running this algorithm in parallel, in order to obtain independent samples from which to obtain the covariance estimate $\widehat{\Sigma}$. For those η_is requiring a Metropolis update, the author suggests using a normal proposal density with variance 1, a logical choice if the orthogonalizing transformation has indeed been successful (though recall Gelman, Roberts, and Gilks (1996) would use variance $(2.4)^2$ instead). Finally, he suggests updating the $\widehat{\Sigma}$ estimate at regular intervals using the current iterates, again in order to improve the orthogonalization and further speed convergence.

While Müller's approach performs quite well in many settings, the orthogonalizing transformation may be impractical, especially in high dimensions where an enormous number of samples will be necessary in order to obtain a good estimate. Moreover, the continual updating of $\widehat{\Sigma}$ amounts to yet another form of hybrid MCMC algorithm: one that adaptively switches the transition kernel based on the output produced by the algorithm so far. While this seems perfectly sensible in this context, adaptive switching does not in general produce a convergent algorithm. This is shown by the following example, which was suggested by G.O. Roberts and developed more fully by Gelfand and Satin (1994).

Example 5.9 Considerthe following simple joint posterior distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\mathbf{0}\,,\, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Suppose we use the following adaptive algorithm. At iteration t, if $X^{(t)} < 0$ we transform to the new parameters $U = X + Y$ and $V = X - Y$, run one Gibbs step under this new (U, V) parametrization, and finally transform back to (X, Y). Elementary multivariate normal calculations show that the joint distribution of (U, V) is also normal with mean vector zero and covariance matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 2(1+\rho) & 0 \\ 0 & 2(1-\rho) \end{pmatrix} .$$

That is, U and V can be independently generated as $N(0, 2(1 + \rho))$ and $N(0, 2(1 - \rho))$ variates, respectively.

If $X^{(t)} > 0$, we do not transform to (U, V)-space, instead drawing from the Gibbs full conditionals in the original parameter space. Clearly both of these transition kernels have the proper stationary distribution, but consider how the output from the adaptive algorithm will look. When $X^{(t)} < 0$, $X^{(t+1)}$ will be positive with probability $1/2$ and negative with probability $1/2$, since its full conditional is symmetric about 0 and is independent of $X^{(t)}$. But when $X^{(t)} > 0$, $X^{(t+1)}$ will be positive with probability greater than $1/2$, since

$$X^{(t+1)} \Big| X^{(t)} \sim N(\rho^2 X^{(t)}\,,\, 1 - \rho^4)\ .$$

Taken together, these two results imply that long sequences from this algorithm will tend to have more positive Xs than negative. This is borne out in the simulation study conducted by Gelfand and Sahu (1994). With $\rho = .9$ and $X^{(0)} \sim N(0,1)$, they obtained $\hat{P}(X^{(100)} > 0) = .784$ and $\hat{E}(X^{(100)}) = .6047$. ■

In summary, adaptive (output-dependent) MCMC algorithms should be used with extreme caution. Perhaps the safest approach is to adapt only in the initial, exploratory phase of the algorithm *(pilot adaptation),* and then stick with a single, nonadaptive approach when saving samples to be used for subsequent inference. Alternatively, recent work by Mira and Sargent (2000) shows that adapting every m iterations for m fixed in advance will in most cases still maintain the chain's stationary distribution. Such *(systematic adaptation)* might be quite useful - say, in improving the current covariance estimate $\widehat{\Sigma}$ in Miiller's approach above, or in the structured MCMC approach described in the next subsection.

While the Gibbs sampler, the Metropolis-Hastings algorithm, and their hybrids are by far the two most common MCMC methods currently in use, there are a variety of extensions and other MCMC algorithms that are extremely useful in analyzing certain specific model settings or achieving particular methodological goals. For example, Bayesian model choice typically requires specialized algorithms, since the models compared are generally not of identical size (see Sections 6.3 and 6.4 for a discussion of these algorithms). The remainder of this subsection describes a few useful specialized algorithms.

Overrelaxation

As already mentioned, high correlations within the sample space can significantly retard the convergence of both Gibbs and univariate Metropolis-Hastings algorithms. Tranformations to reduce correlations, as in Müller's (1991) approach, offer a direct solution but will likely be infeasible in high dimensions. An alternative is to use *overrelaxation*, which attempts to speed the sampled chain's progress through the parameter space by ensuring that each new value in the chain is *negatively* correlated with its predecessor. This method is the MCMC analogue of the overrelaxation methods long used to accelerate iterative componentwise maximization algorithms, such as the Gauss-Seidel method (see e.g. Thisted, 1988, Sec. 3.11.3).

In the case where all full conditional distributions $p(\theta_i|\theta_{j\neq i}, \mathbf{y})$ are Gaussian, Adler (1981) provides an effective overrelaxation method. Suppose θ_i has conditional mean μ_i and conditional variance σ_i^2. Then at iteration t, Adler's method draws $\theta_i^{(t)}$ as

$$\theta_i^{(t)} = \mu_i + \alpha(\theta_i^{(t-1)} - \mu_i) + \sigma_i(1-\alpha^2)^{1/2} Z \, , \qquad (5.28)$$

where $-1 \leq \alpha \leq 1$, and Z is a standard normal random variate. Choosing

$\alpha = 0$ produces ordinary Gibbs sampling; choosing $\alpha < 0$ produces overrelaxation, i.e., a $\theta_i^{(g)}$ sample on the other side of the conditional mean from $\theta_i^{(t-1)}$ ($\alpha > 0$ corresponds to underrelaxation). When $\alpha = -1$, the chain will not be convergent (it will get "stuck" tracing a single contour of the density), so values close to -1 (say, $\alpha = -0.98$) are often recommended.

Several authors generalize this approach beyond Gaussian distributions using transformations or Metropolis accept-reject steps, but these methods are often insufficiently general or suffer from unacceptably high rejection rates. Neal (1998) instead offers a generalization based on order statistics. Here we start by generating K independent values $\{\theta_{i,k}\}_{k=1}^{K}$ from the full conditional $p(\theta_i|\theta_{j\neq i}, \mathbf{y})$. Arranging these values and the old value in nondecreasing order (and breaking any ties randomly), we have

$$\theta_{i,0} \leq \theta_{i,1} \leq \cdots \leq \theta_{i,r} \equiv \theta_i^{(t-1)} \leq \cdots \leq \theta_{i,K} \ , \tag{5.29}$$

so that r is the index of the old value in this list. Then we simply take $\theta_{i,K-r}$ as our new, overrelaxed value for $\theta_i^{(t)}$. The parameter K plays the role of α in Adler's method: $K = 1$ produces Gibbs sampling, while $K \to \infty$ is analogous to choosing $\alpha = -1$ in (5.28).

Neal (1998) shows how to implement this approach without explicit generation of the K random variables in (5.29) by transforming the problem to one of overrelaxation for a $Unif(0,1)$ distribution. This in turn requires the ability to compute the cumulative distribution function and its inverse for each full conditional to be overrelaxed. Still, the method may be worthwhile even in settings where this cannot be done, since the increase in runtime created by the K generations is typically offset by the resulting increase in sampling efficiency. Indeed, the **WinBUGS** package incorporates overrelaxation via direct sampling of the values in (5.29).

Blocking and Structured MCMC

A general approach to accelerating MCMC convergence is *blocking,* i.e., updating multivariate blocks of (typically highly correlated) parameters. Recent work by Liu (1994) and Liu et al. (1994) confirms its good performance for a broad class of models, though Liu et al. (1994, Sec. 5) and Roberts and Satin (1997, Sec. 2.4) give examples where blocking actually slows a sampler's convergence. Unfortunately, success stories in blocking are often application-specific, and general rules have been hard to find.

Sargent, Hodges, and Carlin (2000) introduce a simple, general, and flexible blocked MCMC method for a large class of hierarchical models called *Structured MCMC* (SMCMC). The approach accelerates convergence for many of these models by blocking parameters and taking full advantage of the posterior correlation structure induced by the model and data. As an elementary illustration, consider the balanced one-way random-effects model with *JK* observations y_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, J$. The first stage of the model assumes that $y_{ij} = \theta_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$. At the second stage $\theta_i = \mu + \delta_i$ where $\delta_i \sim N(0, \nu)$, while at the third stage $\mu = M + \xi$ where $\xi \sim N(0, \tau^2)$. Adding prior distributions for σ^2 and ν completes the model specification. Rewriting the second and third stages as

$$\begin{aligned} 0 &= -\theta_i + \mu + \delta_i, \\ \text{and} \quad M &= \mu - \xi \,, \end{aligned}$$

the model can be expressed in the form of a linear model as

$$\left[\begin{array}{c} \mathbf{y} \\ \hline \mathbf{0}_K \\ \hline M \end{array}\right] = \left[\begin{array}{ccc|c} \mathbf{1}_J & \cdots & \mathbf{0}_J & \\ \vdots & \ddots & \vdots & 0_{JK} \\ \mathbf{0}_J & \cdots & \mathbf{1}_J & \\ \hline & -I_{K\times K} & & \mathbf{1}_K \\ \hline & \mathbf{0}'_K & & 1 \end{array}\right] \left[\begin{array}{c} \theta_1 \\ \vdots \\ \theta_K \\ \hline \mu \end{array}\right] + \left[\begin{array}{c} \boldsymbol{\epsilon} \\ \hline \boldsymbol{\delta} \\ \hline -\xi \end{array}\right] \qquad (5.30)$$

where $\mathbf{y}' = (y_{11}, \ldots, y_{1J}, \cdots, y_{K1}, \ldots, y_{KJ})'$, $\boldsymbol{\epsilon}$ is partitioned similarly, and $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_K)'$. Hodges (1998) shows that Bayesian inferences based on (5.30) are identical to those drawn from the standard formulation.

A wide variety of hierarchical models can be reexpressed in a general form of which (5.30) is a special case, namely,

$$\left[\begin{array}{c} \mathbf{y} \\ \hline \mathbf{0} \\ \hline M \end{array}\right] = \left[\begin{array}{c|c} X_1 & \mathbf{0} \\ \hline H_1 & H_2 \\ \hline G_1 & G_2 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\theta}_1 \\ \hline \boldsymbol{\theta}_2 \end{array}\right] + \left[\begin{array}{c} \boldsymbol{\epsilon} \\ \hline \boldsymbol{\delta} \\ \hline \xi \end{array}\right] . \qquad (5.31)$$

In more compact notation, (5.31) can be expressed as

$$Y = X\boldsymbol{\theta} + E, \qquad (5.32)$$

where X and Y are known, $\boldsymbol{\theta}$ is unknown, and E is an error term having $\text{Cov}(E) = \Gamma$ with Γ block diagonal having blocks corresponding to the covariance matrices of each of $\boldsymbol{\epsilon}, \boldsymbol{\delta}$, and ξ, i.e.,

$$\text{Cov}(E) = \Gamma = Diag\,(\text{Cov}(\boldsymbol{\epsilon}), \text{Cov}(\boldsymbol{\delta}), \text{Cov}(\xi)) \ . \qquad (5.33)$$

In our simple example (5.30), Γ is fully (not just block) diagonal.

Hodges (1998) uses the term "data case" to refer to rows of X, Y, and E in (5.32) corresponding to X_1 in (5.31). These are the terms in the joint posterior into which the outcomes $\mathbf{y}$ enter directly. Rows of X, Y, and E corresponding to the H_i are "constraint cases." They place restrictions (stochastic constraints) on possible values of $\boldsymbol{\theta}_1$. Finally, the rows of X, Y, and E corresponding to the G_i are called "prior cases," this label being reserved for cases with known (specified) error variances. We refer to this format as the *constraint case formulation* of a hierarchical model.

For models with normal errors for the data, constraint, and prior cases, Gibbs sampling and (5.32) allow multivariate draws from the marginal posterior of $\boldsymbol{\theta}$. From (5.32), $\boldsymbol{\theta}$ has conditional posterior

$$\boldsymbol{\theta} \mid Y, X, \Gamma \sim N\left((X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}Y\ ,\ (X'\Gamma^{-1}X)^{-1}\right)\ . \tag{5.34}$$

With conjugate priors for the variance components (i.e., gamma or Wishart priors for their inverses), the full conditional for the variance components is a product of gamma and/or Wishart distributions. Convergence for such samplers is often virtually immediate; see Hodges (1998, Sec. 5.2), and the associated discussion by Wakefield (1998). The constraint case formulation works by using, in each MCMC draw, all the information in the model and data about the posterior correlations among the elements of $\boldsymbol{\theta}$.

The algorithm just outlined has three key features. First, it samples $\boldsymbol{\theta}$ (i.e., all of the mean-structure parameters) as a single block. Second, it does so using the conditional posterior covariance of $\boldsymbol{\theta}$, supplied by both the mean structure of the model, captured by X, and the covariance structure of the model, captured by Γ. Finally, it does so using the conditional distribution in (5.34), for suitable definitions of X and Γ. A SMCMC algorithm for a hierarchical model is any algorithm having these three features.

For linear models with normal errors, the suitable values of X and Γ are those in (5.32) and (5.33). These permit several different SMCMC implementations, the simplest of which is the blocked Gibbs sampler discussed above. However, although this Gibbs implementation may converge in few iterations, computing may be slow due to the need to invert a large matrix, $X'\Gamma^{-1}X$, at each iteration. An alternative would be to use (5.34) as a *candidate* distribution in a Hastings independence chain algorithm. Here we might occasionally update Γ during the algorithm's pre-convergence "burn in" period. This *pilot adaptation* scheme (Gilks et al., 1998) is simple but forces us to use a single (possibly imperfect) Γ for the post-convergence samples that are summarized for posterior inference. An even better alternative might thus be to continually update Γ every m iterations for some preselected m (say, m = 100). Mira and Sargent (2000) show that such a *systematic adaptation* scheme can deliver improvements in convergence without disturbing the stationary distribution of the Markov chain.

Example 5.10 A recent clinical trial (Abrams et al., 1994) compared didanosine (ddI) and zalcitabine (ddC) in HIV-infected patients. This dataset is described in detail in Section 8.1; for now, we take the response variable Y_i for patient i to be the change in CD4 count (for which higher levels indicate a healthier immune system) between baseline and the two-month follow-up, for the K = 367 patients who had both measurements. Three binary predictor variables and their interactions are of interest: treatment group (x_{1i}: 1 = ddC, 0 = ddI), reason for eligibility (x_{2i}: 1 = failed ZDV, 0 = intolerant to ZDV), and baseline Karnofsky score (x_{3i}: 1 = score > 85, 0 = score $\leq$ 85; higher scores indicate better health).

Consider the saturated model for this 2^3 factorial design:

$$Y_i = \theta_0 + x_{1i}\theta_1 + x_{2i}\theta_2 + x_{3i}\theta_3 \\ +x_{1i}x_{2i}\theta_{12} + x_{1i}x_{3i}\theta_{13} + x_{2i}x_{3i}\theta_{23} + x_{1i}x_{2i}x_{3i}\theta_{123} + \epsilon_i \ , \tag{5.35}$$

where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, 367$. We use flat priors for the intercept θ_0 and main effects $(\theta_1, \theta_2, \theta_3)$, but place hierarchical constraints on the two- and three-way interaction terms, namely $\theta_l \sim N(0, \sigma_l^2)$ for l = 12, 13, 23, and 123. This linear model is easily written in the form (5.31). Adopting a vague G(0.0001, 0.0001) prior for $\tau \equiv 1/\sigma^2$ (i.e., having mean 1 and variance 10^4) and independent G(1,1) priors for the $h_l \equiv 1/\sigma_l^2$, l = 12,13, 23,123, completes the specification.

Equation (5.34) yields a SMCMC implementation that alternately samples from the multivariate normal full conditional $p(\boldsymbol{\theta}|\mathbf{h}, \tau, \mathbf{y})$, and the gamma full conditionals $p(\tau|\boldsymbol{\theta}, \mathbf{h}, \mathbf{y})$ and $p(h_l|\boldsymbol{\theta}, \tau, \mathbf{y})$. As previously mentioned, this Gibbs sampler is a SMCMC implementation that updates the candidate Γ at every iteration. We also consider a pilot-adaptive SMCMC implementation, in which we update Γ at iterations 1, . . ., 10, then every 10th iteration until iteration 1000, and then use the value of Γ at iteration 1000 in the "production" run, discarding the 1000 iteration burn-in period. In this chain, $p(\boldsymbol{\theta}|\tilde{\mathbf{h}}, \tilde{\tau}, \mathbf{y})$ is used as a candidate distribution for $\boldsymbol{\theta}$ in a Hastings independence subchain, where $\tilde{\mathbf{h}}$ and $\tilde{\tau}$ are the components of Γ at iteration 1000.

Our model is fully conjugate, so **BUGS** handles it easily. Besides a standard univariate Gibbs sampler, **BUGS** allows only blocking of the fixed effects $(\theta_0, \theta_1, \theta_2, \theta_3)$ into a single vector, for which we specified a multivariate normal prior having near-zero precision. For a comparison based on speed alone, we also ran a univariate Gibbs sampler coded in Fortran.

While we caution against putting too much stock in crude runtimes (which are machine-, compiler-, and programmer-dependent), our runtimes (in seconds) to obtain 5000 post-burn-in samples on an Ultra Sparc 20 workstation were as follows: univariate **BUGS**, 37.4; partially blocked **BUGS**, 72.7; univariate Gibbs implemented in Fortran, 27.1; SMCMC using fully blocked Gibbs, 153.0; and SMCMC with pilot adaptation, 18.5. Pilot-adaptive SMCMC dominates the other three in terms of generation rate in "effective samples" per second (see Subsection 5.4.5 for a precise definition of effective sample size); for $\boldsymbol{\theta}$, the advantage is substantial. The **BUGS** chains are fast but produce highly autocorrelated samples, hurting their effective generation rate. By contrast, SMCMC implemented as a fully blocked Gibbs sampler produces essentially uncorrelated draws, but at the cost of long runtimes because of repeated matrix inversions. ■

Next consider hierarchical nonlinear models, i.e., models in which the outcome $\mathbf{y}$ is not linearly related to the parameters in the mean structure. The data cases of such models do not fit the form (5.31), but the constraint and prior cases can often be written as a linear model. Because of the nonlinearity in the data cases, it is now less straightforward to specify the "suitable definitions of X and Γ" needed for a SMCMC algorithm. To do so, we must supply a linear structure approximating the data's contribution to the posterior. We do this by constructing artificial outcome data $\tilde{y}$ with the property that $E(\tilde{y}|\boldsymbol{\theta}_1)$ is roughly equal to $\boldsymbol{\theta}_1$. Rough equality suffices because we use $\tilde{y}$ only in a Metropolis-Hastings implementation to generate candidate draws for $\boldsymbol{\theta}$. Specifically, for a given $\tilde{y}$ with covariance matrix V, we can write the approximate linear model

$$\left[\begin{array}{c} \tilde{y} \\ \hline 0 \\ \hline M \end{array}\right] \approx \left[\begin{array}{c|c} I & 0 \\ \hline H_1 & H_2 \\ \hline G_1 & G_2 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\theta}_1 \\ \hline \boldsymbol{\theta}_2 \end{array}\right] + \left[\begin{array}{c} \epsilon \\ \hline \delta \\ \hline \xi \end{array}\right], \tag{5.36}$$

where $\epsilon \sim N(0, V)$. Equation (5.36) supplies the necessary X matrix for a SMCMC algorithm, and the necessary Γ uses $\text{cov}(\epsilon) = V$ and the appropriate covariance matrices for δ and ξ in (5.33).

The artificial outcome data $\tilde{y}$ can be supplied by two general strategies. The first makes use of crude parameter estimates from the nonlinear part of the model (without the prior or constraint cases), provided they are available. This might occur when the ratio of data elements to parameters is large and the constraint and prior cases are included solely to induce shrinkage in estimation. We call such estimates "unshrunk estimates," and use them and an estimate of their variance as $\tilde{y}$ and V, respectively.

In certain problems, however, such unshrunk estimates may be unstable, or the model may not be identified without constraint and prior cases. Here one can instead run a simple univariate Metropolis algorithm for a small number of iterations, using prior distributions for the variance components that *insist* on little shrinkage. The posterior mean of $\boldsymbol{\theta}$ approximated by this algorithm is a type of artificial data $\tilde{y}$ which, when used with the constraint and prior cases in (5.36), approximates the nonlinear fit well enough for the present purpose. These "low shrinkage" estimates and corresponding estimates of their variance can then be used as $\tilde{y}$ and V, respectively; see Sargent et al. (2000) for implementational details.

Auxiliary variables and slice sampling

Yet another variant on the basic MCMC approach is to use *auxiliary variables*. In this approach, in order to obtain a sample from a particular stationary distribution $p(x), x \in \mathcal{X}$ with x possibly vector-valued, we enlarge the original state space $\mathcal{X}$ using one or more *auxiliary* (or *latent*) variables. In the spirit of the original data augmentation algorithm, a clever enlargement of the state space can broaden the class of distributions we may sample, and also serve to accelerate convergence to the (marginal) stationary distribution. Use of this idea in MCMC sampling arose in statistical physics (Swendsen and Wang, 1987; Edwards and Sokal, 1988), was brought into the statistical mainstream by Besag and Green (1993), and recently developed further by Neal (1997), Higdon (1998), and Damien et al. (1999).

The basic idea is as follows. Let $p(x)$ be the target distribution, likely known only up to a normalizing constant. A new variable u is introduced having conditional distribution $p(u|x)$, so that the joint distribution of u and x is specified as $p(u, x) = p(x)p(u|x)$. We then construct a Markov chain on $\mathcal{X} \times \mathcal{U}$ by alternately updating x and u via Gibbs sampling or some other method that preserves $p(u, x)$, hence the desired marginal distribution $p(x)$. The conditional of u given x is chosen in a "convenient" way, where this can mean either that the auxiliary variable is introduced to allow better mixing or faster convergence of the process defined on the enlarged state space, or that the resulting algorithm is easier to implement or runs faster in terms of CPU time.

In the *simple slice sampler,* we take $u|x \sim Unif(0, p(x))$. This leads to a uniform joint distribution

$$p(x, u) \propto I_{\{0 \le u \le p(x)\}}(x, u) \tag{5.37}$$

whose marginal density for x is clearly $p(x)$. A Gibbs sampler for this distribution would require only two uniform updates: $u|x \sim Unif(0, p(x))$, and $x|u \sim Unif(x : p(x) \ge u)$. Convenient generation of the latter variate (over the "slice" defined by u) thus requires that $p(x)$ be invertible, either analytically or numerically. Once we have a sample of (x, u) values, we simply disregard the u samples, thus obtaining (after a sufficiently long burn-in period) a sample from the target distribution $p(x)$ via marginalization.

More generally, if $p(x) \propto \pi(x)L(x)$ where $L(x)$ *is* a positive function (as in the usual Bayesian context of an unknown posterior p proportional to a prior π times a likelihood L), then a convenient choice for the conditional distribution of u is $Unif(0, L(x))$. More generally we might factor $p(x)$ as $q(x)L'(x)$ where $L'(x)$ *is* a positive function. This is what is generally referred to as the *slice sampler.* Note that, like the Gibbs sampler, this algorithm does not require the user to specify any tuning constants or proposal densities, and yet, like the Metropolis-Hastings algorithm, it allows sampling in nonconjugate model settings. Indeed, Higdon (1998) shows that the Metropolis algorithm can be viewed as a special case of the slice sampler.

Finally, suppose $p(x) \propto \pi(x) \prod_{i=1}^{n} L_i(x)$. We can use a collection of auxilliary variables $u = (u_1, \ldots, u_n)$, where we take $p(u_i|x)$ to be $Unif(0, L_i(x))$

for all i. Now the joint density is given by

$$p(x, u_1, \ldots, u_n) \propto \pi(x) \prod_{i=1}^{n} I_{\{0 \le u_i \le L_i(x)\}}(x, u_i) \; . \tag{5.38}$$

Clearly the marginal density for x is $p(x)$, as desired. The full conditionals for the u_i are $Unif(0, L_i(x))$ while the full conditional for x is simply $\pi(x)$ restricted to the set $\{x : L_i(x) \ge u_i, \; i = 1, \ldots, n\}$. Damien et al. (1999) show that this product slice sampler can be used to advantage in numerous hierarchical and nonconjugate Bayesian settings where the L_i are readily invertible and the intersection set $\{x : L_i(x) \ge u_i, \; i = 1, \ldots, n\}$ is thus easy to compute. However, Neal (1997) points out that introducing so many auxiliary variables (one for each term in the likelihood) may lead to slow convergence. This author recommends hybrid Monte Carlo methods that suppress the random walk behavior, including the overrelaxed version (Neal, 1998) described above. Mira (1998) also shows clever approaches to efficiently sample the variable of interest given the proposed auxiliary variable(s); rejection sampling from $\pi(x)$ may be used as a last resort.

Mira and Tierney (1997) and Roberts and Rosenthal (1999a) give mild and easily verified regularity conditions that ensure that the slice sampler is geometrically and often even uniformly ergodic. Mira and Tierney (1997) show that given any Metropolis-Hastings independence chain algorithm using some $h(x)$ as proposal distribution, a "corresponding" slice sampler can be designed (by taking $q(x) = h(x)$ in the $p(x) = q(x)L'(x)$ factorization above) that produces estimates with uniformly smaller asymptotic variance (on a sweep-by-sweep basis), and that converges faster to stationarity. Roberts and Rosenthal (1999a) also show that the slice sampler is stochastically monotone, which enables development of some useful, quantitative convergence bounds. All of this suggests that, when the slice sampler can be implemented, it has spectacularly good theoretical convergence properties.

Future work with slice sampling looks to its "perfect sampling" extension (Mira, Møller, and Roberts, 1999; see also our brief discussion of this area on page 176), as well as its application in challenging applied settings beyond those concerned primarily with image restoration, as in Hurn (1997) and Higdon (1998). An exciting recent development in this regard is the polar slice sampler (Roberts and Rosenthal, 1999b), which has excellent convergence properties even for high dimensional target densities.

Finally, we hasten to add that there are many closely related papers that explore the connection between expectation-maximization (EM) and data augmentation methods, borrowing recent developments in the former to obtain speedups in the latter. The general idea is the same as that for the auxiliary variable methods described above: namely, to expand the parameter space in order to achieve algorithms which are both suitable for general usage (i.e., not requiring much fine-tuning) and faster than standard Metropolis or Gibbs approaches. For example, the "conditional augmentation" EM algorithm of Meng and Van Dyk (1997) and the "marginal augmentation" parameter expanded EM (PX-EM) algorithm of Liu et al. (1998) are EM approaches that served to motivate the later, specifically MCMC-oriented augmentation schemes of Meng and Van Dyk (1999) and Liu and Wu (1999), the latter of which the authors refer to as parameter expanded data augmentation (PX-DA).

Summary

Many of the hybrid algorithms presented in this subsection attempt to remedy slow convergence due to high correlations within the parameter space. Examples include MCMC algorithms which employ *blocking* (i.e., updating parameters in medium-dimensional groups. as in SMCMC) and *collapsing* (i.e., generating from partially marginalized distributions), as well as overrelaxation and auxiliary variable methods like the slice sampler. But of course, the problem of estimation in the face of high correlations is not anew one: it has long been the bane of maximization algorithms applied to likelihood surfaces. Such algorithms often use a transformation designed to make the likelihood function more regular (e.g.. in 2-space. a switch from a parametrization wherein the effective support of the likelihood function is oblong to one where the support is more nearly circular). Müller's (1991) "orthogonalizing" transformation applied within his adaptive univariate Metropolis algorithm is one such approach; see also Hills and Smith (1992) in this regard. Within the class of hierarchical linear mixed models. Gelfand et al. (1995) discuss simple hierarchical centering reparametrizations that often enable improved algorithm convergence.

Accelerating the convergence of an MCMC algorithm to its stationary distribution is currently an extremely active area of research. Besides those presented here, other promising ideas include resampling and adaptive switching of the transition kernel (Gelfand and Sahu. 1994) and multichain annealing or tempering (Geyer and Thompson. 1995: Neal, 1996a). Gilks and Roberts (1996) give an overview of these and other acceleration methods. In the remainder of this chapter, we continue with variance estimation and convergence investigation for the basic algorithms, returning to these more specialized ones in subsequent chapters as the need arises.

5.4.5 *Variance estimation*

We now turn to the problem of obtaining estimated variances (equivalently, standard errors) for posterior means obtained from MCMC output. Our summary here is rather brief; the reader is referred to Ripley (1987, Chapter 6) for further details and alternative approaches.

Suppose that for a given parameter λ we have a single long chain of

MCMC samples $\{\lambda^{(t)}\}_{t=1}^{N}$, which for now we assume come from the stationary distribution of the Markov chain (we attack the convergence problem in the next subsection). The simplest estimate of $E(\lambda|\mathbf{y})$ is then given by

$$\hat{E}(\lambda|\mathbf{y}) = \hat{\lambda}_N = \frac{1}{N}\sum_{t=1}^{N}\lambda^{(t)} ,$$

analogous to the estimator given in (5.6) for the case of iid sampling. Continuing this analogy, then, we could attempt to estimate $Var(\hat{\lambda}_N)$ by following the approach of (5.7). That is, we would simply use the sample variance, $s_\lambda^2 = \frac{1}{N-1}\sum_{t=1}^{N}(\lambda^{(t)} - \hat{\lambda}_N)^2$, divided by N, obtaining

$$\widehat{Var}_{iid}(\hat{\lambda}_N) = s_\lambda^2/N = \frac{1}{N(N-1)}\sum_{t=1}^{N}(\lambda^{(t)} - \hat{\lambda}_N)^2 .$$

While this estimate is easy to compute, it would very likely be an underestimate due to positive autocorrelation in the MCMC samples. This problem could be ameliorated somewhat by combining the draws from a collection of initially overdispersed sampling chains. Alternatively, one could simply use the sample variance of the N^{th} iteration output from m independent parallel chains, as in Figure 5.4. But as already mentioned, this approach is horribly wasteful, discarding an appalling $(N - 1)/N$ of the samples. A potentially cheaper but similar alternative would be to subsample a single chain, retaining only every k^{th} sample with k large enough that the retained samples are approximately independent. However, MacEachern and Berliner (1994) give a simple proof using the Cauchy-Schwarz inequality that such systematic subsampling from a stationary Markov chain always increases the variance of sample mean estimators (though more recent work by MacEachern and Peruggia, 2000a, shows that the variance can decrease if the subsampling strategy is tied to the actual updates made). Thus, in the spirit of Fisher, it is better to retain all the samples and use a more sophisticated variance estimate, rather than systematically discard a large portion of them merely to achieve approximate independence.

One such alternative uses the notion of *effective sample size,* or ESS (Kass et al. 1998, p. 99). ESS is defined as

$$ESS = N/\kappa(\lambda) ,$$

where $\kappa(\lambda)$ is the *autocorrelation time* for λ, given by

$$\kappa(\lambda) = 1 + 2\sum_{k=1}^{\infty}\rho_k(\lambda) ,$$

where $\rho_k(\lambda)$ is the autocorrelation at lag k for the parameter of interest λ. We may estimate $\kappa(\lambda)$ using sample autocorrelations estimated from the MCMC chain, cutting off the summation when these drop below, say, 0.1

in magnitude. The variance estimate for $\hat{\lambda}_N$ is then

$$\widehat{Var}_{ESS}(\hat{\lambda}_N) = s_\lambda^2 / ESS(\lambda) = \frac{\kappa(\lambda)}{N(N-1)} \sum_{t=1}^{N} (\lambda^{(t)} - \hat{\lambda}_N)^2 .$$

Note that unless the $\lambda^{(t)}$ are uncorrelated, $\kappa(\lambda) > 1$ and ESS(λ) $< N$, so that $\widehat{Var}_{ESS}(\hat{\lambda}_N) > \widehat{Var}_{iid}(\hat{\lambda}_N)$, in concert with intuition. That is, since we have fewer than N effective samples, we expect some inflation in the variance of our estimate.

A final and somewhat simpler (though also more naive) method of estimating $Var(\hat{\lambda}_N)$ is through *hatching.* Divide our single long run of length N into m successive batches of length k *(i.e.,* $N = mk$). with batch means $B_1, \ldots, B_m$. Clearly $\hat{\lambda}_N = \bar{B} = \frac{1}{m}\sum_{i=1}^m B_i$. We then have the variance estimate

$$\widehat{Var}_{batch}(\hat{\lambda}_N) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (B_i - \hat{\lambda}_N)^2 , \tag{5.39}$$

provided that k is large enough so that the correlation between batches is negligible, and m is large enough to reliably estimate $Var(B_i)$. It is important to verify that the batch means are indeed roughly independent - say, by checking whether the lag 1 autocorrelation of the B_i is less than 0.1. If this is not the case, we must increase k (hence N, unless the current m is already quite large), and repeat the procedure.

Regardless of which of the above estimates $\hat{V}$ is used to approximate $Var(\hat{\lambda}_N)$, a 95% confidence interval for $E(\lambda|\mathbf{y})$ is then given by

$$\hat{\lambda}_N \pm z_{.025}\sqrt{\hat{V}} ,$$

where $z_{.025}$ = 1.96, the upper .025 point of a standard normal distribution. If the hatching method is used with fewer than 30 batches, it is a good idea to replace $z_{.025}$ by $t_{m-1,.025}$, the upper .025 point of a t distribution with m - 1 degrees of freedom.

5.4.6 Convergence monitoring and diagnosis

We conclude our presentation on MCMC algorithms with a discussion of their convergence. Because their output is random and autocorrelated, even the definition of this concept is often misunderstood. When we say that an MCMC algorithm has *converged* at time T, we mean that its output can be safely thought of as coming from the true stationary distribution of the Markov chain for all $t > T$. This definition is not terribly precise since we have not defined "safely," but it is sufficient for practical implementation: no pre-convergence samples should be retained for subsequent inference. Some authors refer to this pre-convergence time as the *burn-in* period.

To provide an idea as to why a properly specified MCMC algorithm may be expected to converge, we consider the following simple bivariate two-compartment model, originally analyzed by Casella and George (1992).

Example 5.11 Suppose the joint distribution of the two variables X and Y is summarized by the 2 x 2 layout in Table 5.5. In this table, all $p_i \geq 0$

		X 0	1
Y	0	p_1	p_2
	1	p_3	p_4

Table 5.5 2 x 2 *multinomial distribution of X and Y.*

and $\sum_{i=1}^{4} p_i = 1$. Notating the joint distribution as $f_{x,y}$, we note it can be characterized completely by the 2 x 2 matrix

$$\begin{pmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{pmatrix} = \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix} .$$

The true marginal distribution of X is then

$$f_x = \begin{pmatrix} f_x(0) \\ f_x(1) \end{pmatrix} = \begin{pmatrix} p_1 + p_3 \\ p_2 + p_4 \end{pmatrix} , \tag{5.40}$$

and similarly for Y. The two complete conditional distributions are also easily written in matrix form, namely,

$$A_{y|x} = \begin{pmatrix} f_{y|x}(0|0) & f_{y|x}(1|0) \\ f_{y|x}(0|1) & f_{y|x}(1|1) \end{pmatrix} = \begin{pmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{pmatrix} \tag{5.41}$$

and

$$A_{x|y} = \begin{pmatrix} f_{x|y}(0|0) & f_{x|y}(1|0) \\ f_{x|y}(0|1) & f_{x|y}(1|1) \end{pmatrix} = \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix} . \tag{5.42}$$

$A_{y|x}$ and $A_{x|y}$ can be thought of as *transition matrices,* giving the probabilities of getting from the two X-states to the two Y-states, and vice versa. In fact, if all we care about is the X-marginal $f_x(x)$, we could consider the X sequence alone, since it is also a Markov chain with transition matrix

$$A_{x|x} = A_{y|x}A_{x|y} ,$$

which gives $P\left(X^{(t+1)} = x|X^{(t)} = x\right)$. Repeating this process t times, we have that $(A_{x|x})^t$ gives $P\left(X^{(t)} = x|X^{(0)} = x\right)$. Writing $f_x^{(t)}$ as the distribution of $X^{(t)}$, we have

$$\begin{aligned} f_x^{(t)} &= f_x^{(0)}(A_{x|x})^t \\ &= f_x^{(0)}(A_{x|x})^{t-1} \cdot A_{x|x} \\ &= f_x^{(t-1)}A_{x|x} . \end{aligned}$$

Provided that every entry in $A_{x|x}$ is positive, a fundamental theorem from stochastic processes assures that for any $f_x^{(0)}$, $f_x^{(t)}$ converges to a vector f_x which satisfies

$$f_x = f_x A_{x|x} \ . \tag{5.43}$$

(A check of this fact in our scenario is left as Exercise 19.) Hence using any starting distribution $f_x^{(0)}$, $f_x^{(t)}$ does converge to f_x as $t \to \infty$, with $f_x^{(t)}$ being a better and better approximation as t grows larger. ■

Using a rigorous mathematical framework, many authors have attempted to establish conditions for convergence of various MCMC algorithms in broad classes of problems. For example, Roberts and Smith (1993) provide relatively simple conditions for the convergence of the Gibbs sampler and the Metropolis-Hastings algorithm. Roberts and Tweedie (1996) show the geometric ergodicity of a broad class of Metropolis-Hastings algorithms, which in turn provides a central limit theorem (i.e., asymptotic normality of suitably standardized ergodic sums of the output from such an algorithm); see also Meyn and Tweedie (1993, Chapter 17) in this regard. Results such as these require elements of advanced probability theory that are well beyond the scope of this book, and hence we do not discuss them further. We do however discuss some of the common causes of convergence failure, and provide a brief review of several diagnostic tools used to make stopping decisions for MCMC algorithms.

Overparametrization and identifiability

Ironically, the most common source of MCMC convergence difficulties is a result of the methodology's own power. The MCMC approach is so generally applicable and easy to use that the class of candidate models for a given dataset now appears limited only by the user's imagination. However, with this generality has come the temptation to fit models so large that their parameters are *unidentified*, or nearly so. To see why this translates into convergence failure, consider the problem of finding the posterior distribution of θ_1 where the likelihood is defined by

$$Y_i|\theta_1, \theta_2 \stackrel{iid}{\sim} N(\theta_1 + \theta_2 \ , \ 1) \ , \ \ i = 1, \ldots, n \ , \tag{5.44}$$

and we adopt independent flat priors for both θ_1 and θ_2. Only the sum of the two parameters is identified by the data, so without proper priors for θ_1 and θ_2, their marginal posterior distributions will be improper as well. Unfortunately, a naive application of the Gibbs sampler in this setting would not reveal this problem. The complete conditionals for θ_1 and θ_2 are both readily available for sampling as normal distributions. And for any starting point, the sampler would remain reasonably stable, due to the ridge in the likelihood surface $L(\theta_1, \theta_2)$ at $\theta_1 + \theta_2 = \bar{y}$. The inexperienced user might be tempted to use a smoothed histogram of the $\theta_1^{(g)}$ samples obtained as a (proper) estimate of the (improper) posterior density $p(\theta_1|\bar{y})$!

An experienced analyst might well argue that Gibbs samplers like the one described in the preceding paragraph are perfectly legitimate, *provided* that their samples are used only to summarize the posterior distributions of identifiable functions of the parameters (in this case, $g(\boldsymbol{\theta}) = \theta_1 + \theta_2$). But while the deficiency in model (5.44) is immediately apparent, in more complicated settings (e.g., hierarchical and random effects models) failures in identifiability can be very subtle. Moreover, models that are overparametrized (either deliberately or accidentally) typically lead to high posterior correlations among the parameters (*cross correlations*) which will dramatically retard the movement of the Gibbs sampler through the parameter space. Even in models which *are* identified, but "just barely" so (e.g., model (5.44) with a vague but proper prior for θ_1), such high crosscorrelations (and the associated high *autocorrelations* in the realized sample chains) can lead to excruciatingly slow convergence. MCMC algorithms defined on such spaces are thus appropriate only when the model permits a firm understanding of which parametric functions are well-identified and which are not.

For the underidentified Gaussian linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with X less than full column rank, Gelfand and Sahu (1999) provide a surprising MCMC convergence result. They show that under a flat prior on $\boldsymbol{\beta}$, the Gibbs sampler for the full parameter vector $\boldsymbol{\beta}$ is divergent, but the samples from the identified subset of parameters (say, $\boldsymbol{\delta} = X_1\boldsymbol{\beta}$) form an *exact* sample from their (unique) posterior density $p(\boldsymbol{\delta}|\mathbf{y})$. That is, such a sampler will produce identically distributed draws from the true posterior for $\boldsymbol{\delta}$, and convergence is immediate. In subsequent work, Gelfand, Carlin, and Trevisam (2000) consider a broad class of Gaussian models with covariates, namely,

$$\mathbf{Y} = (X_0 \;\; X_0\Delta_1 \;\; \cdots \;\; X_0\Delta_r)\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r)'$. It then turns out $Corr(\boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t+1)})$ approaches 0 as the prior variance component for $\boldsymbol{\beta}_0$ goes to infinity once the chain has converged. This then permits independent sampling from the posterior distributions of estimable parameters. Exact sampling for these parameters is also possible in this case provided that *all* of the $\boldsymbol{\beta}$ prior variance components go to infinity. Simulation work by these authors suggests these results still hold under priors (rather than fixed values) for the variance components that are imprecise but with large means. For more on overparametrization and posterior impropriety, see Natarajan and McCulloch (1995) Hobert and Casella (1996), and Natarajan and Kass (2000). Carlin and Louis (2000) point out that these dangers might motivate a return to EB analysis in such cases; c.f. Ten Have and Localio (1999) in this regard.

We have already mentioned how a transition matrix (or kernel, for a continuous parameter space) can be analyzed to conclude whether or not the corresponding MCMC algorithm will converge. But in order to apply the method in actual practice, we need to know not only if the algorithm will converge, but when. This is a much more difficult problem. Theoretical results often provide convergence rates which can assist in selecting between competing algorithms (e.g., linear versus quadratic convergence), but since the rates are typically available only up to an arbitrary constant, they are of little use in deciding when to stop a given algorithm. Recently, however, some authors have made progress in obtaining bounds on the number of iterations T needed to guarantee that distribution being sampled at that time, $p^{(T)}(\boldsymbol{\theta})$, is in some sense within ϵ of the true stationary distribution, $p(\boldsymbol{\theta})$. For example, Polson (1996) develops polynomial time convergence bounds for a discrete jump Metropolis algorithm operating on a log-concave target distribution in a discretized parameter space. Rosenthal (1993, 1995a,b) instead uses Markov minorization conditions, providing bounds in continuous settings involving finite sample spaces and certain hierarchical models. Approaches like these hold great promise, but typically involve sophisticated mathematics in sometimes laborious derivations. Cowles and Rosenthal (1998) ameliorate this problem somewhat by showing how auxiliary simulations may often be used to verify the necessary conditions numerically, and at the same time provide specific values for use in the bound calculations. A second difficulty is that the bounds obtained in many of the examples analyzed to date in this way are fairly loose, suggesting numbers of iterations that are several orders of magnitude beyond what would be reasonable or even feasible in practice (though see Rosenthal, 1996, for some tight bounds in model (3.9), the two-stage normal-normal compound sampling model).

A closely related area that is the subject of intense recent research is *exact* or *perfect sampling*. This refers to MCMC simulation methods that can guarantee that a sample drawn at a given time will be exactly distributed according to the chain's stationary distribution. The most popular of these is the "coupling from the past" algorithm, initially outlined for discrete state spaces by Propp and Wilson (1996). Here, a collection of chains from different initial states are run at a sequence of starting times going *backward* into the past. When we go far enough back that all the chains have "coupled" by time 0, this sample is guaranteed to be an exact draw from the target distribution. Green and Murdoch (1999) extend this approach to continuous state spaces. The idea has obvious appeal, since it seems to eliminate the convergence problem altogether! But to date such algorithms have been made practical only for relatively small problems within fairly well-defined model classes; their extension to high-dimensional models (in which the conditioning offered by the Gibbs sampler would be critical) remains an open problem. Moreover, coupling from the past delivers only a *single* exact sample; to obtain the collection of samples needed for inference, the coupling algorithm would need to be repeated a large number of times. This would produce uncorrelated draws, but at considerably more expense than the usual "forward" MCMC approach, where if we can be sure we are past the burn-in period, every subsequent (correlated) sample can safely be thought of as a draw from the true posterior.

As a result of the practical difficulties associated with theoretical convergence bounds and exact sampling strategies, almost all of the applied work involving MCMC methods has relied on the application of diagnostic tools to output produced by the algorithm. Early attempts by statisticians in this regard involved comparing the empirical distributions of output produced at consecutive (or nearly consecutive) iterations, and concluding convergence when the difference between the two was negligible in some sense. This led to samplers employing a large number of parallel, independent chains, in order to obtain simple moment, quantile, and density estimates.

Example 5.12 Recall the rat population growth model from Example 5.6. Gelfand et al. (1990) used $m = 50$ parallel replications, and judged sampler convergence by monitoring density estimates of the form (5.16) every 5 iterations (far enough apart to be considered independent). They concluded convergence at iteration $T = 35$ based on a "thick felt-tip pen test": the density estimates plotted at this iteration differed by less than the width of a thick felt-tip pen. Setting aside concerns about the substantial wastage of samples in this highly parallel approach (see Figure 5.4), not to mention the establishment of an international standard for MCMC convergence assessment pens, this diagnostic is unsuitable for another reason: it is easily fooled into prematurely concluding convergence in slowly mixing samplers. This is because it measures the distance separating the sampled distribution at two different iterations, rather than that separating either from the true stationary distribution. ■

Of course, since the stationary distribution will always be unknown to us in practice, this same basic difficulty will plague *any* convergence diagnostic. Indeed, this is what leads many theoreticians to conclude that all such diagnostics are fundamentally unsound. In the context of the previous example, we would like to ensure that the distance between the true distribution and our estimate of it is small, i.e. , $\int |\hat{f}_t(x) - f(x)| dx < \epsilon$. But unfortunately, by the triangle inequality we have

$$\int \left|\hat{f}_t(x) - f(x)\right| dx \leq \int \left|\hat{f}_t(x) - f_t(x)\right| dx + \int |f_t(x) - f(x)|\, dx \, . \quad (5.45)$$

The first term in the sum, sometimes referred to as the "mean squared error component" or "Monte Carlo noise," becomes small as m $\to \infty$, by the Law of Large Numbers. The second term, sometimes referred to as the "bias component" or "convergence noise," becomes small as $t \to \infty$, by the convergence of the Markov chain to its stationary distribution. Convergence diagnosis is thus concerned with choosing T so that the second term is sufficiently small; m can then be increased to whatever level is required to make the first term small as well. Note that for a fixed generation budget, there could be a tradeoff in the choice of t and m. But in any event, all we can hope to see is $\int |\hat{f}_t(x) - \hat{f}_{t+k}(x)|dx$, which is not directly related to any of the three quantities in (5.45).

Despite the fact that no diagnostic can "prove" convergence of a MCMC algorithm (since it uses only a finite realization from the chain), many statisticians rely heavily on such diagnostics, if for no other reason than "a weak diagnostic is better than no diagnostic at all." We agree that they do have value, though it lies primarily in giving a rough idea as to the nature of the mixing in the chain, and perhaps in keeping the analyst from making a truly embarrassing mistake (like reporting results from an improper posterior). As such, we now present a summary of a few of the methods proposed, along with their goals, advantages, and limitations. Cowles and Carlin (1996) and Mengersen et al. (1999) present much more complete reviews, describing the methodological basis and practical implementation of many such diagnostics that have appeared in the recent literature. Brooks and Roberts (1998) offer a similar review focusing on the underlying mathematics of the various approaches, rather than their implementation.

Convergence diagnostics

MCMC convergence diagnostics have a variety of characteristics, which we now list along with a brief explanation.

1. *Theoretical basis.* Convergence diagnostics have been derived using a broad array of mathematical machinery and sophistication, from simple cusum plots to kernel density estimates to advanced probability theory. Less sophisticated methods are often more generally applicable and easier to program, but may also be more misleading in complex modeling scenarios.
2. *Diagnostic goal.* Most diagnostics address the issue of *bias,* or the distance of the estimated quantities of interest at a particular iteration from their true values under the target distribution. But a few also consider *variance,* or the precision of those estimates, which is perhaps more important from a practical standpoint.
3. *Output format.* Some diagnostics are quantitative, producing a single number summary, while others are qualitative, summarized by a graph or other display. The former type is easier to interpret (and perhaps even automate), but the latter can often convey additional, otherwise hidden information to the experienced user.

4. *Replication.* Some diagnostics may be implemented using only a single MCMC chain, while others require a (typically small) number of parallel chains. (This is one of the more controversial points in this area, as we shall illustrate below.)

5. *Dimensionality.* Some diagnostics consider only univariate summaries of the posterior, while others attempt to diagnose convergence of the full joint posterior distribution. Strictly speaking, the latter type is preferable, since posterior summaries of an apparently converged parameter may be altered by slow convergence of other parameters in the model. Unfortunately, joint posterior diagnostics are notoriously difficult and likely infeasible in high-dimensional problems, leading us to substitute univariate diagnostics on a representative collection of parameters.

6. *Applicability.* Some diagnostics apply only to output from Gibbs samplers, some to output from any MCMC scheme, and some to a subset of algorithms somewhere between these two extremes. Again, there is often a tradeoff between broad applicability and diagnostic power or interpretability.

7. *Ease of use.* Finally, generic computer code is freely available to implement some convergence diagnostics. Generic code may be written by the user for others, and subsequently applied to the output from any MCMC sampler. Other diagnostics require that problem-specific code be written, while still others require advance analytical work in addition to problem-specific code. Not surprisingly, the diagnostics that come with their own generic code are the most popular among practitioners.

We now consider several diagnostics, highlighting where they lie with respect to the above characteristics. Perhaps the single most popular approach is due to Gelman and Rubin (1992). Here, we run a small number (m) of parallel chains with different starting points, as in Figures 5.6, 5.7, and 5.8. These chains must be initially overdispersed with respect to the true posterior; the authors recommend use of a preliminary mode-finding algorithm to assist in this decision. Running the m chains for $2N$ iterations each, we then attempt to check whether the variation within the chains for a given parameter of interest λ approximately equals the total variation across the chains during the latter N iterations. Specifically, we monitor convergence by the estimated *scale reduction factor,*

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{N-1}{N} + \frac{m+1}{mN}\frac{B}{W}\right)\frac{df}{df-2}}\,, \qquad (5.46)$$

where B/N is the variance between the means from the m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of an approximating t density to the posterior distribution. Equation (5.46) is the factor by which the scale parameter of the t density might shrink if sampling were continued indefinitely; the authors show it must approach 1 as $N \to \infty$. (See Appendix Subsection C.3.3 for information on how to obtain an S implementation of this algorithm off the web at no charge.)

Gelman and Rubin's approach can be derived using standard statistical techniques, and is applicable to output from any MCMC algorithm. Thanks to the authors' S function, it is quite easy to use and produces a single summary number on an easily interpreted scale (i.e., values close to 1 suggest good convergence). However, it has been criticized on several counts. First, since it is a univariate diagnostic, it must be applied to each parameter or at least a representative subset of the parameters; the authors also suggest thinking of minus twice the log of the joint posterior density as an overall summary parameter to monitor. The approach focuses solely on the bias component of convergence, providing no information as to the accuracy of the resulting posterior estimates, i.e.,

$$\hat{E}(\lambda|\mathbf{y}) = \frac{1}{mN} \sum_{j=1}^{m} \sum_{t=N+1}^{2N} \lambda_j^{(t)} .$$

Finally, the method relies heavily on the user's ability to find a starting distribution that is actually overdispersed with respect to the true posterior of λ, a condition we can't really check without knowledge of the latter!

Brooks and Gelman (1998) extend the Gelman and Rubin approach three important ways, describing an iterated graphical implementation of the original approach, broadening its definition to avoid the original's implicit normality assumption, and developing a multivariate generalization for simultaneous convergence diagnosis for every parameter in a model (although the practical usefulness of this final extension is somewhat open to question). The diagnostics remain fairly easy to compute, though unfortunately there is no generally available S code available for its implementation.

In a companion paper to the original Gelman and Rubin article, Geyer (1992) criticizes the authors' "initial overdispersion" assumption as unverifiable, and the discarding of early samples from several parallel chains as needlessly wasteful. He instead recommends a dramatically different approach wherein we run a *single* chain, and focus not on the bias but on the variance of resulting estimates. For example, letting $\hat{\lambda}_N = \frac{1}{N} \sum_{t=1}^{N} \lambda^{(t)}$ we have the central limit theorem

$$\sqrt{N}(\hat{\lambda}_N - \lambda) \xrightarrow{D} N(0, \sigma^2) ,$$

where σ^2 can be estimated by

$$\hat{\sigma}_N^2 = \sum_{t=-\infty}^{\infty} w_N(t)\, \hat{\gamma}_{N,t} .$$

Here, $w_N(t)$ is a weight function called a lag *window*, and $\hat{\gamma}_{N,t}$ is an estimate of the lag t autocovariance function, namely,

$$\hat{\gamma}_{N,t} = \frac{1}{N} \sum_{i=1}^{N-t} \left(\lambda^{(i)} - \hat{\lambda}_N \right) \left(\lambda^{(i+t)} - \hat{\lambda}_N \right) .$$

The lag window satisfies $0 \leq w_N(t) \leq 1$, and is used to downweight the large-lag terms. An approximate 95% confidence interval for. $E(\lambda|\mathbf{y})$ is thus obtainable as

$$\hat{\lambda}_N \pm 1.96 \hat{\sigma}_N / \sqrt{N} . \quad (5.47)$$

Note that simply using the standard deviation of the $\lambda^{(t)}$ samples in place of $\hat{\sigma}_N$ in (5.47) would be anti conservative, since the sampling chain will likely feature positive autocorrelation, leading to an underestimate of σ. Note that the above formulae assume no discarding of early samples; Geyer argues that less than 1% of the run will normally be a sufficient burn-in period whenever the run is long enough to give much precision (recall Gelman and Rubin discard the first 50% of each chain).

Geyer's approach relies on more sophisticated mathematics, addresses the variance goal, and avoids parallel chains or any notion of "initial overdispersion." However, it doesn't generalize well beyond posterior moments (e.g., to posterior quantiles or density estimates for λ), and does not concern itself with actually looking for bias, simply starting its sole chain at the MLE or some other "likely" place. Finally, it is less intuitive and a bit unwieldy, since proper choice of the lag window is not obvious.

The diagnostic of Raftery and Lewis (1992) begins by retaining only every k^{th} sample after burn-in, with k large enough that the retained samples are approximately independent. More specifically, their procedure supposes our goal is to estimate a posterior quantile $q = P(\lambda < u|\mathbf{y})$ to within $\pm$r units with probability s. (That is, with $q = .025$, $r = .005$, and $s = .95$, our reported 95% credible sets would have true coverage between .94 and .96.) To accomplish this, they assume $Z_t^{(k)} = Z_{1+(t-1)k}$ is a Markov chain, where $Z_t = I(\lambda^{(t)} < u)$, and use results from Markov chain convergence theory.

The Raftery and Lewis approach addresses both the bias and variance diagnostic goals, is broadly applicable, and is relatively easy to use thanks to computer code provided by the authors (see Appendix Subsection C.3.3 for information on how to obtain it). However, Raftery and Lewis' formulation in terms of a specific posterior quantile leads to the rather odd result that the time until convergence for a given parameter differs for different quantiles. Moreover, systematic subsampling typically inflates the variance of the resulting estimates, as pointed out in Subsection 5.4.5.

New MCMC convergence diagnostics continue to appear; see e.g. the variety of clever ideas in Robert (1995). Another notable recent addition to the literature is that of Johnson (1998), who derives a scheme based on the times at which the chain "couples" with an auxiliary chain that is occasionally restarted from a fixed point in the parameter space. The approach uses the idea of *regeneration* times, points in the algorithm that divide the Markov chain into sections whose sample paths are independent; see Mykland, Tierney and Yu (1995).

Summary

Cowles and Carlin (1996) compare the performance of several convergence diagnostics in two relatively simple models, and conclude that while some are often successful, all can also fail to detect the sorts of convergence failures they were designed to identify. As a result, for a generic application of an MCMC method, they recommend a variety of diagnostic tools, rather than any single plot or statistic. Here, then, is a possible diagnostic strategy:

- Run a few (3 to 5) parallel chains, with starting points drawn (perhaps systematically, rather than at random) from a distribution believed to be overdispersed with respect to the stationary distribution (say, covering ± 3 prior standard deviations from the prior mean).
- Visually inspect these chains by overlaying their sampled values on a common graph (as in Figures 5.6 , 5.7, and 5.8 for each parameter, or, for very high-dimensional models, a representative subset of the parameters. (For a standard hierarchical model, such a subset might include most of the fixed effects, some of the variance components. and a few well-chosen random effects - say, corresponding to two individuals who are at opposite extremes relative to the population.)
- Annotate each graph with the Gelman and Rubin (1992) statistic and lag 1 autocorrelations, since they are easily calculated and the latter helps to interpret the former (large G&R statistics may arise from either slow mixing or multimodality).
- Investigate crosscorrelations among parameters suspected of being confounded, just as one might do regarding collinearity in linear regression.

The use of multiple algorithms with a particular model is often helpful, since each will have its own convergence properties and may reveal different features of the posterior surface. Indeed, for low-dimensional problems, recent adaptive versions of traditional quadrature and importance sampling methods often serve as good checks and may even be more efficient; see e.g. Evans and Swartz (1995) in this regard. Considering multiple models is also generally a good idea, since besides deepening our understanding of the data this improves our chances of detecting convergence failure or hidden problems with model identifiability.

Note that convergence *acceleration* is a task closely related to, but more tractable than, convergence *diagnosis.* In many problems, the time spent in a dubious convergence diagnosis effort might well be better spent searching for ways to accelerate the MCMC algorithm, since if more nearly uncorrelated samples can be obtained, the diagnosis problem may vanish. We thus remind the reader that reparametrizations can often improve a model's correlation structure, and hence speed convergence: see Hills and Smith (1992) for a general discussion and Gelfand et al. (1995, 1996) for "hierarchical centering" approaches for random effects models. The more sophisticated MCMC algorithms presented in Subsection 5.4.4 (mixtures and cycles, overrelaxation, structured MCMC, slice sampling, etc.) also often lead to substantial reductions in runtime.

Finally, since automated convergence monitoring, even using a variety of diagnostics, is often unsafe, we instead recommend a two-stage process, wherein model specification and associated sampling are separated from convergence diagnosis and subsequent output analysis. In this regard, Best, Cowles, and Vines (1995) have developed a collection of *S*+ routines called **CODA** for the second stage of this process, with the first stage being accomplished by **BUGS**, the recently-developed software package for Bayesian analysis using Gibbs sampling (Gilks et al., 1992). Appendix Subsection C.3.3 provides a discussion of these programs, guidance on how to use them, and instructions on how to obtain them.

5.5 Exercises

1. Use equation (5.1) to obtain a first-order approximation to the Bayes factor *BF*, given in equation (2.20). Is there any computational advantage in the case where the two models are nested, i.e., $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_2, \boldsymbol{\gamma})$?

2. Suppose that $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$. Describe how equation (5.5) could be used to investigate the effect of single case deletion (i.e., removing x_i from the dataset, for $i = 1, \ldots, n$) on the posterior expectation of $g(\boldsymbol{\theta})$.

3. Show that the probability of accepting a given θ_j candidate in the rejection sampling is c/M, where $c = \int L(\theta)\pi(\theta)d\theta$, the normalizing constant for the posterior distribution $h(\theta)$.

4. Suppose we have a sample $\{\theta_1, \ldots, \theta_N\}$ from a posterior $p_1(\theta|\mathbf{y})$, and we want to see the effect of changing the prior from $\pi_1(\theta)$ to $\pi_2(\theta)$. Describe how the weighted bootstrap procedure could be helpful in this regard. Are there any restrictions on the shape of the new prior π_2 for this method to work well?

5. Use quadratic regression to analyze the data in Table 5.6, which relate property taxes and age for a random sample of 19 homes. The response variable Y is property taxes assessed on each home (in dollars), while the explanatory variable X is age of the home (in years). Assume a normal likelihood and the noninformative prior $\pi(\boldsymbol{\beta}, \sigma) = 1/\sigma$, where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2)$ and the model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$. Find the

Y	X	Y	X
925	1	480	20
870	2	486	22
809	4	462	25
720	4	441	25
694	5	426	30
630	8	368	35
626	10	350	40
562	10	348	50
546	12	322	50
523	15		

Table 5.6 County property tax (Y) and age of home (X)

first two moments (hence a point and interval estimate) for the intercept β_0 assuming:

(a) standard frequentist methods,

(b) asymptotic posterior normality (first order approximation), and

(c) Laplace's method (second order approximation).

6. In the previous problem, suppose that instead of the intercept β_0, the parameter of interest is the age corresponding to minimum tax under this model. Describe how you would modify the above analysis in this case.

7. Actually perform the point and interval estimation of the age corresponding to minimum property tax in the previous problem using a Monte Carlo, instead of an asymptotic, approach. That is, generate $\boldsymbol{\beta}^{(i)}$ values from the joint posterior distribution, and then, using the formula for the minimum tax age in terms of the βs, obtain Monte Carlo values from its posterior distribution.

   (Hint: $\sigma^2|\mathbf{y} \sim IG(\frac{n-p}{2}, \frac{2}{RSS})$, where IG denotes the inverse gamma distribution, $p = 3$, the number of regressors, and $RSS = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2$, the residual sum of squares.)

   Do you think the quadratic model is appropriate for these data? Why or why not?

8. Consider the following hierarchical changepoint model for the number of occurrences Y_i of some event during time interval is

$$Y_i \sim \begin{cases} \text{Poisson}(\theta), & i = 1, \ldots, k \\ \text{Poisson}(\lambda), & i = k+1, \ldots, n \end{cases}$$

Year	Count	Year	Count	Year	Count	Year	Count
1851	4	1879	3	1907	0	1935	2
1852	5	1880	4	1908	3	1936	1
1853	4	1881	2	1909	2	1937	1
1854	1	1882	5	1910	2	1938	1
1855	0	1883	2	1911	0	1939	1
1856	4	1884	2	1912	1	1940	2
1857	3	1885	3	1913	1	1941	4
1858	4	1886	4	1914	1	1942	2
1859	0	1887	2	1915	0	1943	0
1860	6	1888	1	1916	1	1944	0
1861	3	1889	3	1917	0	1945	0
1862	3	1890	2	1918	1	1946	1
1863	4	1891	2	1919	0	1947	4
1864	0	1892	1	1920	0	1948	0
1865	2	1893	1	1921	0	1949	0
1866	6	1894	1	1922	2	1950	0
1867	3	1895	1	1923	1	1951	1
1868	3	1896	3	1924	0	1952	0
1869	5	1897	0	1925	0	1953	0
1870	4	1898	0	1926	0	1954	0
1871	5	1899	1	1927	1	1955	0
1872	3	1900	0	1928	1	1956	0
1873	1	1901	1	1929	0	1957	1
1874	4	1902	1	1930	2	1958	0
1875	4	1903	0	1931	3	1959	0
1876	1	1904	0	1932	3	1960	1
1877	5	1905	3	1933	1	1961	0
1878	5	1906	1	1934	1	1962	1

Table 5.7 *Number of British coal mining disasters by year*

$$\theta \sim G(a_1, b_1),\ \lambda \sim G(a_2, b_2),\ \theta, \text{ and } \lambda \text{ independent;}$$
$$b_1 \sim IG(c_1, d_1),\ b_2 \sim IG(c_2, d_2),\ b_1 \text{ and } b_2 \text{ independent,}$$

where G denotes the gamma and IG the inverse gamma distributions. Apply this model to the data in Table 5.7, which gives counts of coal mining disasters in Great Britain by year from 1851 to 1962. (Here, "disaster" is defined as an accident resulting in the deaths of 10 or more miners.) Set $a_1 = a_2 = .5$, $c_1 = c_2 = 1$, and $d_1 = d_2 = 1$ (a collection of "moderately informative" values). Also assume $k = 40$ (corresponding to the year 1890), and obtain marginal posterior density estimates for

θ, λ, and $R = \theta/\lambda$ using formula (5.16) with output from the Gibbs sampler.

9. In the previous problem, assume k is unknown, and adopt the following prior:

$$k \sim \text{Discrete Uniform}(1, \ldots, n), \text{ independent of } \theta \text{ and } \lambda \ .$$

Add k into the sampling chain, and obtain a marginal posterior density estimate for it. What is the effect on the posterior for R?

10. In the previous problem, replace the third-stage prior given above with

$$b_1 \sim G(c_1, d_1),\ b_2 \sim G(c_2, d_2),\ b_1 \text{ and } b_2 \text{ independent},$$

thus destroying the conjugacy for these two complete conditionals. Resort to rejection or Metropolis-Hastings subsampling for these two components instead, choosing appropriate values for c_1, c_2, d_1, and d_2. What is the effect on the posterior for b_1 and b_2? For R? For k?

11. Reanalyze the flour beetle mortality data in Table 5.4, replacing the multivariate Metropolis algorithm used in Example 5.7 with

(a) a Hastings algorithm employing independence chains drawn from a $N(\tilde{\boldsymbol{\theta}}, \tilde{\Sigma})$ candidate density, where $\tilde{\boldsymbol{\theta}}$ = (1.8, -4.0, -1.0)' (roughly the true posterior mode) and $\tilde{\Sigma}$ as given in (5.27).

(b) a univariate Metropolis (Metropolis within Gibbs) algorithm using proposal densities $N(\theta_i^{(t-1)}, D_{ii})$, i = 1, 2, 3, with D as given in equation (5.26).

(c) a cycle or mixture of these two algorithms, as suggested in Example 5.8.

Based on your results, what are the advantages and disadvantages of each approach?

12. Returning again to the flour beetle mortality data and model of Example 5.7, note that the decision to use $\tilde{\Sigma} = 2\hat{\Sigma}$ in equation (5.27) was rather arbitrary. That is, univariate Metropolis "folklore" suggests a proposal density having variance roughly twice that of the true target should perform well, creating bigger jumps around the parameter space while still mimicking the target's correlation structure. But the *optimal* amount of variance inflation might well depend on the dimension and precise nature of the target distribution, the type of sampler used (multivariate versus univariate, Metropolis versus Hastings, etc.), or any number of other factors.

Explore these issues in the context of the flour beetle mortality data in Table 5.4 by resetting $\tilde{\Sigma} = c\hat{\Sigma}$ for c = 1 (candidate variance matched to the target) and c = 4 (candidate standard deviations twice those in the target) using

(a) the multivariate Hastings algorithm in part (a) of problem 11.

(b) the univariate Metropolis algorithm in part (b) of problem 11.

(c) the multivariate Metropolis algorithm originally used in Example 5.7.

Evaluate and compare performance using acceptance rates and lag 1 sample autocorrelations. Do your results offer any additions (or corrections) to the "folklore?"

13. Spiegelhalter et al. (1995b) analyze the flour beetle mortality data in Table 5.4 using the **BUGS** language. These authors use only the usual, two-parameter parametrization for $p_i \equiv P(death|w_i)$, but compare three different link functions,

$$
\begin{aligned}
\text{logit:} \quad p_i &= \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} , \\
\text{probit:} \quad p_i &= \Phi(\alpha + \beta z_i) , \\
\text{complementary log-log:} \quad p_i &= 1 - \exp[-\exp(\alpha + \beta z_i)] ,
\end{aligned}
$$

where we use the centered covariate $z_i = w_i - \bar{w}$ in order to reduce correlation between the intercept α and the slope β. Using the Windows-based version of this package, **WinBUGS**, the code to implement this model is as follows:

```
model

  for (i in 1:k){
    y [i] - dbin(p[i], n [i])
    logit(p[i]) <- alpha + beta*(w[i]-mean(w[]))
#    probit(p[i]) <- alpha + beta*(w[i]-mean(w[]))
#    cloglog(p[i]) <- alpha + beta*(w[i]-mean(w[]))

} # end of i loop

  alpha " dnorm(0.0, 1.0E-3)
  beta ' dnorm(0.0, 1.0E-3)

} # end of program
```

Anything after a "#" is a comment in **WinBUGS**, so this version corresponds to the logit model. Also, "dbin" denotes the binomial distribution and "dnorm" the normal distribution, where in the latter case the second argument gives the normal *precision,* not variance (thus the normal priors for α and β have precision 0.001, approximating flat priors).

(a) The full conditional distributions for α and β have no closed form, but **WinBUGS** does recognize them as being log-concave, and thus capable of being sampled using the Gilks and Wild (1992) adaptive rejection algorithm. Prove this log-concavity under the logit model.

(b) Actually carry out the data analysis in **WinBUGS**, after downloading theprogram from http://www.mrc-bsu.cam.ac.uk/bugs/ .(The program above, along with properly formatted data and initial value lists, are included in the "examples" section of the help materials.) Do the estimated posteriors for the dosage effect β substantially differ for different link functions?

14. Consider again the model of Example 5.6. Suppose that instead of a straight-line growth curve, we wish to investigate the *exponential* growth model,

$$Y_{ij} \overset{ind}{\sim} N\left(\alpha_i \exp(\beta_i x_{ij})\,,\, \sigma^2\right),\; i = 1, \ldots, k,\; j = 1, \ldots, n_i\;,$$

with the remaining priors and hyperpriors as previously specified.

(a) Of the full conditional distributions for $\{\alpha_i\}, \{\beta_i\}, \alpha_0, \beta_0, \Sigma^{-1}$ and σ^2, which now require modification?

(b) Of those distributions in your previous answer, which may still be derived in closed form as members of familiar families? Obtain expressions for these distributions.

(c) Of those full conditionals that lack a closed form, give a Metropolis or Hastings substep for generating the necessary MCMC samples.

(d) Write a program applying your results to the data in Table 5.3, and obtain estimated posterior distributions for α_0 and β_0. What does your fitted model suggest about the growth patterns of young rats?

15. Consider the generalized Hastings algorithm which uses the following candidate density:

$$q(v, u) = \begin{cases} p(v_i | u_{j\neq i}) & \text{for } v_{j\neq i} = u_{j\neq i} \\ 0 & \text{otherwise} \end{cases} .$$

That is, the algorithm chooses (randomly or deterministically) an index $i \in \{1, \ldots, K\}$, and then uses the full conditional distribution along the i^{th} coordinate as the candidate density. Show that this algorithm has Hastings acceptance ratio (5.25) identically equal to 1 (as called for by the Gibbs algorithm), and hence that the Gibbs sampler is a special case of the Hastings algorithm.

16. Show that Adler's overrelaxation method (5.28) leaves the desired distribution invariant. That is, show that if $\theta_i^{(g-1)}$ is conditionally distributed as $N(\mu_i, \sigma_i^2)$, then so is $\theta_i^{(g)}$.

17. A random variable Z defined on (0, ∞ is said to have a *D-distribution* with parameters $\delta, \beta > 0$ and $k \in$ 10,1,2.... } if its density function is defined (up to a constant of proportionality) by

$$p_Z(z) \propto z^{\delta-1}e^{-\beta z}(1-e^{-z})^k \ . \tag{5.48}$$

This density emerges in many Bayesian nonparametric problems (e.g., Damien, Laud, and Smith, 1995). We wish to generate observations from this density.

(a) If Z's density is log-concave, we know we can generate the necessary samples using the method of Gilks and Wild (1992). Find a condition (or conditions) under which this density is guaranteed to be log-concave.

(b) When Z's density is not log-concave, we can instead use an auxiliary variable approach (Walker, 1995), adding the new random variable U = $(U_1, \ldots, U_k)$. The U_i are defined on (0, 1) and mutually independent given Z, such that the joint density function of Z and U is defined (up to a constant of proportionality) by

$$p_{Z,U}(z,u) \propto z^{\delta-1}e^{-\beta z}\prod_{i=1}^{k} I_{(e^{-z},1)}(u_i) \ .$$

Show that this joint density function $p_{Z,U}$ does indeed have marginal density function p_Z given in equation (5.48) above.

(c) Find the full conditional distributions for Z and U, and describe a sampling algorithm to obtain the necessary Z samples. What special subroutines would be needed to implement your algorithm?

18.(a) Consider two approaches for sampling from a N(0, 1) target density, $p(x) \propto \exp(-\frac{1}{2}x^2)$:

- a standard Metropolis algorithm using a Gaussian proposal density, $x^*|x^{(t-1)} \sim N(x^{(t-1)}, \sigma^2)$, and
- the simple slice sampler (5.37).

Run single chains of 1000 iterations for each of these samplers, starting both chains at $x^{(0)} = 0$. For which values of the Metropolis proposal variance σ^2 does the slice sampler produce smaller lag 1 autocorrelations (i.e., faster convergence)?

(b) Repeat the preceding investigation for a *multivariate* $N_5(\mathbf{0}, I_5)$ target density, $p(\mathbf{x}) \propto \exp(-\frac{1}{2}\mathbf{x}'\mathbf{x})$, where $\mathbf{x}' = (x_1, \ldots, x_5)$. We now compare

- a univariate Metropolis algorithm, updating one component at a time, and
- a product slice sampler (5.38), which introduces five auxiliary variables $(u_1, \ldots, u_5)$.

Now which sampler performs better? Is the efficiency of slice sampling affected by the dimension increase?

19. Using the transition matrices (5.41) and (5.42), check that (5.43) holds, i.e., that the true marginal density f_x given in (5.40) is indeed a stationary point of the Gibbs sampler in this problem.

20. Consider the following two complete conditional distributions, originally analyzed by Casella and George (1992):

$$f(x|y) \propto ye^{-yx}, \quad 0 < x < B < \infty$$
$$f(y|x) \propto xe^{-xy}, \quad 0 < y < B < \infty$$

 (a) Obtain an estimate of the marginal distribution of X when $B = 10$ using the Gibbs sampler.

 (b) Now suppose $B = \infty$ so that the complete conditional distributions are ordinary (untruncated) exponential distributions. Show *analytically* that $f_x(t) = 1/t$ is a solution to the integral equation

$$f_x(x) = \int \left[\int f_{x|y}(x|y) f_{y|t}(y|t) dy \right] f_x(t) dt$$

 in this case. Would a Gibbs sampler converge to this solution? Why or why not?

21. Consider the balanced, additive, one-way ANOVA model.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \qquad (5.49)$$

where $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, $\mu \in \Re$, $\alpha_i \in \Re$, and $\sigma_e^2 > 0$. We adopt a prior structure that is a product of independent conjugate priors. wherein μ has a flat prior, $\alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha^2)$, and $\sigma_e^2 \sim IG(a, b)$. Assume that σ_α^2, a, and b are known.

 (a) Derive the full conditional distributions for μ, α_i, and σ_e^2, necessary for implementing the Gibbs sampler in this problem.

 (b) What is meant by "convergence diagnosis?" Describe some tools you might use to assist in this regard. What might you do to improve a sampler suffering from "slow convergence?"

 (c) Suppose for simplicity that σ_e^2 is also known. What conditions on the data or the priors might lead to slow convergence for μ and the (*Hint:* What conditions weaken the identifiability of the parameters?)

 (d) Now let $\eta_i = \mu + \alpha_i$, so that η_i "centers" α_i. Then we can consider two possible parametrizations: (1) μ-α, and (2) μ-η. Generate a sample of data from likelihood (5.49), assuming $I = 5$, $J = 1$, and $\sigma_e = \sigma_\alpha = 1$. Investigate the sample crosscorrelations and autocorrelations produced by Gibbs samplers operating on Parametrizations 1 and 2 above. Which performs better?

(e) Rerun your program for the case where $\sigma_e = 1$ and $\sigma_\alpha = 10$. Now which parametrization performs better? What does this suggest about the benefits of hierarchical centering reparametrizations?

22. To further study the relationship between identifiability and MCMC convergence, consider again the two-parameter likelihood model

$$Y \sim N(\theta_1 + \theta_2 \ , \ 1) \ ,$$

with prior distributions $\theta_1 \sim N(a_1, b_1^2)$ and $\theta_2 \sim N(a_2, b_2^2)$, θ_1 and θ_2 independent.

(a) Clearly θ_1 and θ_2 are individually identified only by the prior; the likelihood provides information only on $\mu = \theta_1 + \theta_2$. Still, the full conditional distributions $p(\theta_1|\theta_2, y)$ and $p(\theta_2|\theta_1, y)$ are available as normal distributions, thus defining a Gibbs sampler for this problem. Find these two distributions.

(b) In this simple problem we can also obtain the marginal posterior distributions $p(\theta_1|y)$ and $p(\theta_2|y)$ in closed form. Find these two distributions. Do the data update the prior distributions for these parameters?

(c) Set $a_1 = a_2 = 50$, $b_1 = b_2 = 1000$, and suppose we observe $y = 0$. Run the Gibbs sampler defined in part (a) for $t = 100$ iterations, starting each of your sampling chains near the prior mean (say, between 40 and 60), and monitoring the progress of θ_1, θ_2, and μ. Does this algorithm "converge" in any sense? Estimate the posterior mean of μ. Does your answer change using $t = 1000$ iterations?

(d) Now keep the same values for a_1 and a_2 but set $b_1 = b_2 = 10$. Again run 100 iterations using the same starting values as in part (b). What is the effect on convergence? Again repeat your analysis using 1000 iterations; is your estimate of $E(\mu|y)$ unchanged?

(e) Summarize your findings, and make recommendations for running and monitoring convergence of samplers running on "partially unidentified" and "nearly partially unidentified" parameter spaces.

CHAPTER 6

Model criticism and selection

To this point we have seen the basic elements of Bayes and empirical Bayes methodology, arguments on behalf of their use from several different philosophical standpoints, and an assortment of computational algorithms for carrying out the analysis. We have observed that the generality of the methodology coupled with the power of modern computing enables consideration of a wide variety of hierarchical models for a given dataset. Given all this, the most natural questions for the reader to ask might be:

1. How can I tell if any of the assumptions I have made (e.g., the specific choice of prior distribution) is having an undue impact on my results?

2. How can I tell if my model is providing adequate fit to the data?

3. Which model (or models) should I ultimately choose for the final presentation of my results?

The first question concerns the *robustness* of the model, the second involves *assessment* of the model, and the third deals with *selection* of a model (or group of models). An enormous amount has been written on these three subjects over the last fifty or so years, since they are the same issues faced by classical applied statisticians. The Bayesian literature on these subjects is of course smaller, but still surprisingly large, especially given that truly applied Bayesian work is a relatively recent phenomenon. We group the three areas together here since they all involve criticism of a model that has already been fit to the data.

Subsection 2.3.3 presented the fundamental tool of Bayesian model selection, the Bayes factor, while Section 2.4 outlined the basics of model assessment and improved estimation and prediction through model averaging. Armed with the computational techniques presented in Chapter 5, we now revisit and expand on these model building tools. While we cannot hope to review all of what has been done, in this chapter we will attempt to present the tools most useful for the applied Bayesian, along with sufficient exemplification so that the reader may employ the approaches independently. As much of the work in this area is problem-specific and/or fairly technical, in certain areas our discussion will be limited to a summary of the technique and a reference to the relevant place in the literature for more detailed information.

6.1 Bayesian robustness

As we have already mentioned, a commonly voiced concern with Bayesian methods is their dependence on various aspects of the modeling process. Possible sources of uncertainty include the prior distribution, the precise form of the likelihood, and the number of levels in the hierarchical model. Of course, these are concerns for the frequentist as well. For example, the statistics literature is filled with methods for investigating the effect of case deletion, which is a special type of likelihood sensitivity. Still, the problem appears more acute for the Bayesian due to the appearance of the prior distribution (even though it may have been deliberately chosen to be noninformative).

In the next subsection, we investigate the robustness of the conclusions of a Bayesian analysis by checking whether or not they remain essentially unchanged in the presence of perturbations in the prior. likelihood. or some other aspect of the model. Subsection 6.1.2 considers a reverse attack on the problem, wherein we attempt to characterize the class of model characteristics (say, prior distributions) that lead to a certain conclusion given the observed data. Both of these approaches differ slightly from what many authors call the "robust Bayesian viewpoint," which seeks model settings (particularly prior distributions) that are robust from the outset. rather than picking a (perhaps informative) baseline and checking it later. or cataloging assumptions based on a predetermined decision. The case studies in Chapter 8 feature a variety of robustness investigations. while further methodological tools are provided in Chapter 7.

6.1.1 Sensitivity analysis

The most basic tool for investigating model uncertainty is the *sensitivity analysis.* That is, we simply make reasonable modifications to the assumption in question, recompute the posterior quantities of interest. and see whether they have changed in a way that has practical impact on interpretations or decisions. If not, the data are strongly informative with respect to this assumption and we need not worry about it further. If so. then our results are sensitive to this assumption, and we may wish to communicate the sensitivity, think more carefully about it, collect more data. or all of these.

Example 6.1 Suppose our likelihood features a mean parameter θ and a variance parameter σ^2 and in our initial analysis we employed a $N(\mu, \tau^2)$ prior on θ and an IG(a, b) prior on σ^2. To investigate possible sensitivity of our results to the prior for θ we might first consider shifts in its mean, e.g., by increasing and decreasing μ by one prior standard deviation τ. We also might alter its precision, say, by doubling and halving τ. Alternatively, we could switch to a heavier-tailed t prior with the same moments.

Sensitivity to the prior for σ^2 could be investigated in the same way, although we would now need to solve for the hyperparameter values required to increase and decrease its mean or double and halve its standard deviation. ■

The increased ease of computing posterior summaries via Monte Carlo methods has led to a corresponding increased ease in performing sensitivity analyses. For example, given a converged MCMC sampler for a posterior distribution p(θ|y), it should take little time for the sampler to adjust to the new posterior distribution $p_{NEW}(\theta|\mathbf{y})$ arising from some modest change to the prior or likelihood. Thus, the additional sampling could be achieved with a substantial reduction in the burn-in period. Still, for a model having p hyperparameters, simply considering two alternatives (one larger, one smaller) to the baseline value for each results in a total of 3^p possible priors - possibly infeasible even using MCMC methods.

Fortunately, a little algebraic work eliminates the need for further sampling. Suppose we have a sample $\theta_1, \ldots, \theta_N\}$ from a posterior p(θ|y), which arises from a likelihood $f(\mathbf{y}|\theta) = \prod_{i=1}^n f(y_i|\theta)$ and a prior p(θ). To study the impact of deleting case k, we see that the new posterior is

$$p_{NEW}(\theta|\mathbf{y}) \propto \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(y_k|\theta)} \propto \frac{1}{f(y_k|\theta)} p(\theta|\mathbf{y}) .$$

In the notation of Subsection 5.3.2, we can use $g(\theta) = p(\theta|\mathbf{y})$ as an importance sampling density, so that the weight function is given by

$$w(\theta) = \frac{p_{NEW}(\theta|\mathbf{y})}{p(\theta|\mathbf{y})} = \frac{1}{f(y_k|\theta)} .$$

Thus for any posterior function of interest $h(\theta)$, we have $\hat{E}[h(\theta)|\mathbf{y}] = \sum_{j=1}^N h(\theta_j)/N$ and

$$\hat{E}_{NEW}[h(\theta)|\mathbf{y}] = \frac{\sum_{j=1}^N h(\theta_j)w(\theta_j)}{\sum_{j=1}^N w(\theta_j)} , \tag{6.1}$$

using equation (5.8). Since p and p_{NEW} should be reasonably similar, the former should be a good importance sampling density for the latter and hence the approximation in equation (6.1) should be good for moderate N. We could also obtain an approximate sample from p_{NEW} via the weighted bootstrap by resamplina θ_i values with probability q_i, where $q_i = w(\theta_i)/\sum_{j=1}^N w(\theta_j)$.

In situations where large changes in the prior are investigated, or where outlying or influential cases are deleted, p_{NEW} may differ substantially from p. A substantial difference between the two posteriors can lead to highly variable importance sampling weights and, hence, poor estimation. Importance link function transformations of the Monte Carlo sample (MacEachern and Peruggia, 2000b) can be used to effectively alter the distribution from which the $\theta_1, \ldots, \theta_N\}$ were drawn, stablizing the importance sampling weights and improving estimation.

Sensitivity analysis via asymptotic approximation

Notice that importance sampling may be used for sensitivity analysis even if the original posterior sample $\{\theta_1, \ldots, \theta_N\}$ was obtained using some other method (e.g., the Gibbs sampler). The new posterior estimates in (6.1) are not likely to be as accurate as the original, but are probably sufficient for the purpose of a sensitivity analysis. In the same vein. the asymptotic methods of Section 5.2 can be used to obtain approximate sensitivity results without any further sampling or summation. For example, we have already seen in equation (5.5) a formula requiring only two maximizations (both with respect to the original posterior) that enables any number of sensitivity investigations. In the realm of model comparison. Kass and Vaidyanathan (1992) use such approximations to advantage in determining the sensitivity of Bayes factors to prior and likelihood input. Let $\pi_{NEW,k}(\theta_k)$ be new priors on θ_k where $k = 1, 2$ indexes the model. Writing BF for the Bayes factor under the original priors and BF_{NEW} for the Bayes factor under the new priors, Kass and Vaidyanathan (1992, equation (2.23)) show that

$$BF_{NEW} = BF \cdot \frac{r_1(\tilde{\theta}_1)}{r_2(\tilde{\theta}_2)} \cdot \left\{1 + O\left(\frac{1}{n}\right)\right\} \tag{6.2}$$

where $r_k(\tilde{\theta}_k) = \pi_{NEW,k}(\tilde{\theta}_k)/\pi_k(\tilde{\theta}_k)$ and $\tilde{\theta}_k$ is the posterior mode using the original prior $\pi_k(\theta_k)$. Thus, one only needs to evaluate r_k at $\tilde{\theta}_k$ for k=1, 2 for each new pair of priors, and these computations are sufficiently easy and rapid that a very large number of new priors can be examined without much dificulty.

A further simplification comes from a decomposition of the parameter vector into two components, $\theta_k = (\theta_k^{(1)}, \theta_k^{(2)})$ with the first component containing all parameters that are common to both models. We define $\theta^{(1)} = \theta_1^{(1)}$ $(= \theta_2^{(1)})$. This parameter vector might be considered a nuisance parameter while the $\theta_k^{(2)}$ $k = 1, 2$ become the parameter vectors of interest. If we take the priors to be such that the two components are a priori independent under both models, we may write $\pi_k(\theta_k) = \pi_k^{(1)}(\theta^{(1)}) \cdot \pi_k^{(2)}(\theta_k^{(2)})$. Suppose this is the case, and furthermore $\pi_1^{(1)}(\theta^{(1)}) = \pi_2^{(1)}(\theta^{(1)})$, and that both these conditions hold for the new priors considered. In this situation, if it turns out that the first-component posterior modes under the two models are approximately equal (formally, to order $O(n^{-1})$), then

$$\frac{\pi_{NEW,1}(\tilde{\theta}_1)}{\pi_{NEW,2}(\tilde{\theta}_2)} \doteq \frac{\pi^{(2)}_{NEW,1}(\tilde{\theta}^{(2)}_1)}{\pi^{(2)}_{NEW,2}(\tilde{\theta}^{(2)}_2)}$$

to the same order of accuracy as equation (6.2). The ratio $r_1(\tilde{\theta}_1)/r_2(\tilde{\theta}_2)$ in (6.2) thus involves the new priors only through their second components. In other words, if the modal components $\tilde{\theta}^{(1)}_k$ are nearly equal for $k = 1$ and 2, when we perform the sensitivity analysis we do not have to worry about the effect of modest modifications to the prior on the nuisance-parameter component of θ_k. Carlin, Kass, Lerch, and Huguenard (1992) use this approach for investigating sensitivity of a comparison of two competing models of human working memory load to outliers and changes in the prior distribution.

Sensitivity analysis via scale mixtures of normals

The conditioning feature of MCMC computational methods enables another approach to investigating the sensitivity of distributional specifications in either the likelihood or prior for a broad class of common hierarchical models. Consider the model $y_i = \mu_i + \epsilon_i, i = 1, \ldots, n,$ where the μ_i are unknown mean structures and the ϵ_i are independent random errors having density f with mean 0. For convenience, one often assumes that the ϵ_i form a series of independent normal errors, i.e., $\epsilon_i|\sigma^2 \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. However, the normal distribution's light tails may well make this an unrealistic assumption, and so we wish to investigate alternative forms that allow greater variability in the observed y_i values. Andrews and Mallows (1974) show that expanding our model to

$$\epsilon_i|\sigma^2, \lambda_i \sim N(0, \lambda_i\sigma^2),\ i = 1, \ldots, n\ ,$$

and subsequently placing a prior on λ_i enables a variety of familiar (and more widely dispersed) error densities to emerge. That is, we create f as the *scale mixture of normal distributions,*

$$f(\epsilon_i|\sigma^2) = \int_\Lambda p(\epsilon_i|\sigma^2, \lambda_i)p(\lambda_i)d\lambda_i\ ,\ i = 1, \ldots, n\ .$$

The following list identifies the necessary functional forms for $p(\lambda_i)$ to obtain some of the possible departures from normality:

- Student's t errors: If $\nu/\lambda_i \sim \chi^2_\nu$ (i.e., if $\lambda_i \sim IG(\frac{\nu}{2}, \frac{2}{\nu})$) then $\epsilon_i|\sigma \sim t_\nu(0, \sigma)$.

- Double exponential errors: If $\lambda_i \sim E(2)$, the exponential distribution having mean 2, then $\epsilon_i|\sigma \sim DE(0, \sigma)$, where DE denotes the double exponential distribution.

- Logistic errors: If $1/\sqrt{\lambda_i}$ has the asymptotic Kolmogorov distance distribution, then $\epsilon_i|\sigma$ is logistic (see Andrews and Mallows, 1974).

Generally one would not view the addition of n parameters to the model as a simplifying device, but would prefer instead to work directly with the nonnormal error density in question. However, Carlin and Polson (1991) point out that the conditioning feature of the Gibbs sampler makes this augmentation of the parameter space quite natural. To see this, note first that under an independence prior on the λ_i, we have $\lambda_i|\{\lambda_{j\neq i}\}, \{\mu_j\}, \sigma^2, \mathbf{y} \sim \lambda_i|\mu_i, \sigma^2, y_i$. But by Bayes theorem, $p(\lambda_i|\mu_i, \sigma^2, y_i) \propto p(y_i|\mu_i, \sigma^2, \lambda_i)p(\lambda_i)$, where the appropriate normalization constant, $p(y_i|\mu_i, \sigma^2)$, is known by construction as the desired nonnormal error density. Hence the complete conditional for λ_i will always be of known functional form. Generation of the required samples may be done directly if this form is a standard density; otherwise, a carefully selected rejection method may be employed. Finally, since the remaining complete conditionals (for σ^2 and the μ_j are determined given $\boldsymbol{\lambda}$, convenient prior specifications may often be employed with the normal likelihood, again leading to direct sampling.

Example 6.2 Consider the model $y_i = \mu_i + \epsilon_{ij}$, where j indexes the model error distribution (i.e., the mixing distribution for the λ_i), and i indexes the observation, as before. Suppose that $\mu_i = f(\mathbf{x}_i, \theta)\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a k-dimensional vector of linear nuisance parameters, and $f(\mathbf{x}_i, \theta) = (f_1(\mathbf{x}_i, \theta), \ldots, f_k(\mathbf{x}_i, \theta))$ is a collection of known functions, possibly nonlinear in θ. Creating the $n \times k$ matrix $F_\theta = (f(\mathbf{x}_i, \theta))$, the log-likelihood $\log p(\mathbf{y} \mid \theta, \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2)$ is

$$-\frac{1}{2\sigma^2}(\mathbf{y} - F_\theta\boldsymbol{\beta})^T\Sigma_i^{-1}(\mathbf{y} - F_\theta\boldsymbol{\beta}) - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n}\log\lambda_i \ ,$$

where $\Sigma_i = Diag(\lambda_1, \ldots, \lambda_n)$. Assume that $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \Sigma_0)$, where $\boldsymbol{\beta}_0$ and Σ_0 are known. In addition, let $\sigma^2 \sim IG(a_0, b_0)$, and let θ have prior distribution $p(\theta)$.

Suppose that, due to uncertainty about the error density and the impact of possible outliers, we wish to compare the three models

$$M = 1: \quad \epsilon_i \sim N(0, \sigma^2) \ ,$$

$$M = 2: \quad \epsilon_i \sim t(0, \sigma^2, \nu = 2) \ , \text{ and}$$

$$M = 3: \quad \epsilon_i \sim DE(0, \sigma) \ ,$$

where M indicates which error distribution (model) we have selected. For $M = 1$, clearly the λ_i are not needed, so their full conditional distributions are degenerate at 1. For $M = 2$, we have the full conditional distribution

$$IG\left(\frac{\nu+1}{2}, \left\{\frac{1}{2}\left[\frac{(y_i - f(\mathbf{x}_i, \theta)\boldsymbol{\beta})^2}{\sigma^2} + \nu\right]\right\}^{-1}\right), \ i = 1, \ldots, n \ ,$$

in a manner very similar to that of Andrews and Mallows (1974). Finally, for M = 3 we have a complete conditional proportional to

$$\lambda_i^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\lambda_i + \frac{(y_i - f(\mathbf{x}_i, \theta)\boldsymbol{\beta})^2}{\lambda_i \sigma^2}\right)\right) ,$$

that is, $\lambda_i|\theta, \boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim GIG\left(\frac{1}{2}, 1, (y_i - f(\mathbf{x}_i, \theta)\boldsymbol{\beta})^2/\sigma^2\right)$ for $i = 1, \ldots, n$, where *GIG* denotes the generalized inverse Gaussian distribution (see Devroye, 1986, p. 478). In order to sample from this density, we note that it is the reciprocal of an

$$\text{Inverse Gaussian}\left(\left|\frac{\sigma}{y_i - f(\mathbf{x}_i, \theta)\boldsymbol{\beta}}\right|, 1\right) ,$$

a density from which we may easily sample (see for example Devroye, 1986, p. 149). ■

Thus, the scale mixture of normals approach enables investigation of nonnormal error distributions as a component of the original model, rather than later as part of a sensitivity analysis, and with little additional computational complexity. If the λ_i all have posterior distributions centered tightly about 1, the additional modeling flexibility is unnecessary, and the assumption of normality is appropriate. In this same vein, the λ_i can be thought of as outlier diagnostics, since extreme observations will correspond to extreme fitted values of these scale parameters. Wakefield et al. (1994) use this idea to detect outlying individuals in a longitudinal study by employing the normal scale mixture at the *second* stage of their hierarchical model (i.e., on the prior for the random effect component of μ_i).

6.1.2 Prior Partitioning

The method of sensitivity analysis described in the previous subsection is a direct and conceptually simple way of measuring the effect of the prior distributions and other assumptions made in a Bayes or empirical Bayes analysis. However, it does not free us from careful development of the original prior, which must still be regarded as a reasonable baseline. This can be impractical, since the prior beliefs and vested interests of the potential consumers of our analysis may be unimaginably broad. For example, if we are analyzing the outcome of a government-sponsored clinical trial to determine the effectiveness of a new drug relative to a standard treatment for a given disease, our results will likely be read by doctors in clinical practice, epidemiologists, regulatory workers (e.g., from the U.S. Food and Drug Administration), legislators, and, of course, patients suffering from or at risk of contracting the disease itself. These groups are likely to have widely divergent opinions as to what constitutes "reasonable" prior opinion: the clinician who developed the drug is likely to be optimistic about its value, while the regulatory worker (who has seen many similar drugs emerge as ineffective) may be more skeptical. What we need is a method for communicating the robustness of our conclusions to *any* prior input the reader deems appropriate.

A potential solution to this problem is to "work the problem backward," as follows. Suppose that, rather than fix the prior and compute the posterior distribution, we fix the posterior (or set of posteriors) that produce a given conclusion, and determine which prior inputs are consistent with this desired result, given the observed data. The reader would then be free to determine whether the outcome was reasonable according to whether the prior class that produced it was consistent with his or her own prior beliefs. We refer to this approach simply as *prior partitioning* since we are subdividing the prior class based on possible outcomes, though it is important to remember that such partitions also depend on the data and the decision to be reached. As such, the approach is not strictly Bayesian (the data are playing a role in determining the prior), but it does provide valuable robustness information while retaining the framework's philosophical, structural, documentary, and communicational advantages.

To illustrate the basic idea, consider the point null testing scenario $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Without loss of generality, set $\theta_0 = 0$. Suppose our data x has density $f(x|\theta)$, where θ is an unknown scalar parameter. Let π represent the prior probability of H_0, and $G(\theta)$ the prior cumulative distribution function (cdf) of θ conditional on $\{\theta \neq 0\}$. The complete prior cdf for θ is then $F(\theta) = \pi I_{[0,\infty)}(\theta) + (1-\pi)G(\theta)$, where Is is the indicator function of the set S. Hence the posterior probability of the null hypothesis is

$$P_G(\theta = 0|x) = \frac{\pi f(x|0)}{\pi f(x|0) + (1-\pi)\int f(x|\theta)dG(\theta)} . \qquad (6.3)$$

Prior partitioning seeks to characterize the G for which this probability is less than or equal to some small probability p $\in$ (0, 1), in which case we reject the null hypothesis. (Similarly, we could also seek the *G* leading to $P_G(\theta \neq 0|x) \leq p$, in which we would reject H_1. Elementary calculations show that characterizing this class of priors $\{G\}$ is equivalent to characterizing the set $\mathcal{H}_c$, defined as

$$\mathcal{H}_c = \left\{ G : \int f(x|\theta)dG(\theta) \geq c = \frac{1-p}{p}\frac{\pi}{1-\pi}f(x|0) \right\} . \qquad (6.4)$$

Carlin and Louis (1996) establish results regarding the features of $\mathcal{H}_c$, and then use these results to obtain sufficient conditions for $\mathcal{H}_c$ to be nonempty for classes of priors that satisfy various moment and percentile restrictions. The latter are somewhat more useful, since percentiles and tail areas of the conditional prior *G* are transform-equivariant, and Chaloner et al. (1993) have found that elicitees are most comfortable describing their opinions through a "best guess" (mean, median, or mode) and a few relatively extreme percentiles (say, the 5^{th} and the 95^{th}).

To lay out the technical detail of this percentile restriction approach, let θ_L and θ_U be such that $P_G(\theta \leq \theta_L) = a_L$ and $P_G(\theta > \theta_U) = a_U$, where a_L and a_U lie in the unit interval and sum to less than 1. For fixed values for θ_L and θ_U we seek the region of a_L, a_U values (i.e., the class of priors) that lead to a given decision. To accomplish this we can take a general G, alter it to pass through θ_L, a_L and $\theta_U, 1 - a_U)$, and then search over all G subject only to the constraint that the intervals $(-\infty, a_L]$, $(a_L, a_U]$, and (a_U, ∞) all have positive probability.

Assume that $f(x|\theta)$ is a unimodal function of θ for fixed x that vanishes in both tails, an assumption that will be at least approximately true for large datasets due to the asymptotic normality of the observed likelihood function. Keeping θ_L, θ_U and a fixed, we seek the supremum and infimum of $\int f(x|\theta)dG(\theta)$. The infimum will always be given by

$$(1 - a_L - a_U) \min\{f(x|\theta_L), f(x|\theta_U)\}\ , \tag{6.5}$$

since a unimodal f must take its minimum over the central support interval at one of its edges. However, depending on the location of the maximum likelihood estimator $\hat{\theta}$, $\sup_G \int f(x|\theta)dG(\theta)$ equals

$$\begin{array}{ll} a_L f(x|\hat{\theta}) + (1 - a_L - a_U) f(x|\theta_L) + a_U f(x|\theta_U), & \hat{\theta} \leq \theta_L \\ a_L f(x|\theta_L) + (1 - a_L - a_U) f(x|\hat{\theta}) + a_U f(x|\theta_U), & \theta_L < \hat{\theta} \leq \theta_U \\ a_L f(x|\theta_L) + (1 - a_L - a_U) f(x|\theta_U) + a_U f(x|\hat{\theta}), & \hat{\theta} > \theta_U \end{array} . \tag{6.6}$$

Notice that the infimum is obtained by pushing mass as far away from the MLE as allowed by the constraints, while the supremum is obtained by pushing the mass as close to the MLE as possible. In conjunction with π, the prior probability of H_0, the supremum and infimum may be used to determine the prior percentiles compatible with $P(\theta = 0|x) \leq p$ and $P(\theta \neq 0|x) \leq p$, respectively. Since $\mathcal{H}_c$ is empty if the supremum does not exceed c, we can use the supremum expression to determine whether there are any G that satisfy the inequality in (6.4), i.e., whether any priors G exist that enable stopping to reject the null hypothesis. Similarly, the infimum expression may be useful in determining whether any G enable stopping to reject the alternative hypothesis, H_1.

We may view as fixed either the (θ_L, θ_U) pair, the (a_L, a_U) pair, or both. As an example of the first case, suppose we seek the (a_L, a_U) compatible with a fixed (θ_L, θ_U) pair (an *indifference zone)* for which $\int f(x|\theta)dG(\theta) \geq c$. Then given the location of $\hat{\theta}$ with respect to the indifference zone, equation (6.6) may be easily solved to obtain the half-plane in which acceptable as must lie. When combined with the necessary additional constraints $a_L \geq 0, a_U \geq 0$, and $a_L + a_U \leq 1$, the result is a polygonal region that is easy to graph and interpret. Graphs of acceptable (θ_L, θ_U) pairs for fixed (a_L, a_U) may be obtained similarly, although the solution of equation (6.6) is now more involved and the resulting regions may no longer be compact. We will return to these ideas in the context of our Section 8.2 data example.

Since the ideas behind prior partitioning are somewhat more theoretical and less computational than those driving sensitivity analysis, the approach has been well developed in the literature. Two often-quoted early references are the paper by Edwards, Lindman, and Savage (1963), who explore robust Bayesian methods in the context of psychological models, and the book by Mosteller and Wallace (1964), who discuss bounds on the prior probabilities necessary to choose between two simple hypotheses (authorship of a given disputed *Federalist* paper by either Alexander Hamilton or James Madison).

The subsequent literature in the area is vast; see Berger (1985, Section 4.7) or, more recently, Berger (1994) for a comprehensive review. Here we mention only a few particularly important papers. In the point null setting, Berger and Sellke (1987) and Berger and Delampady (1987) show that we attain the minimum of $P(\theta = \theta_0|x)$ over all conditional priors G for $\theta \neq \theta_0$ when G places all of its mass at $\hat{\theta}$, the maximum likelihood estimate of 0. Even in this case, where G is working with the data against H_0 these authors showed that the resulting $P(\theta = \theta_0|x)$ values are typically still larger than the corresponding two-sided p-value, suggesting that the standard frequentist approach is biased against H_0 in this case. In the interval null hypothesis setting, prior partitioning is reminiscent of the work of O'Hagan and Berger (1988), who obtain bounds on the posterior probability content of each of a collection of intervals that form the support of a univariate parameter, under the restriction that the prior probability assignment to these intervals is in a certain sense unimodal. In the specific context of clinical trial monitoring, Greenhouse and Wasserman (1994) compute bounds on posterior expectations and tail areas (stopping probabilities) over an **e**-contaminated class of prior distributions (Berger and Berliner, 1986).

Further restricting the prior class

Sargent and Carlin (1996) extend the above approach to the case of an interval null hypothesis, i.e., $H_0 : \theta \in [\theta_L, \theta_U]$ versus $H_1 : \theta \notin [\theta_L, \theta_U]$. This formulation is useful in the context of clinical trial monitoring, where $[\theta_L, \theta_U]$ is thought of as an *indifference zone,* within which we are indifferent as to the use of treatment or placebo. For example, we might take $\theta_U > 0$ if there were increased costs or toxicities associated with the treatment. Let π again denote the prior probability of H_0, and let $G(\theta$ now correspond to the prior cdf of θ given $\theta \notin [\theta_L, \theta_U]$. Making the simplifying assumption of a uniform prior over the indifference zone, the complete prior density for θ may be written as

$$p(\theta) = \frac{\pi}{\theta_U - \theta_L} I_{[\theta_L, \theta_U]}(\theta) + (1 - \pi) g(\theta) \ . \qquad (6.7)$$

Sargent and Carlin (1996) derive expressions similar to (6.3) and (6.4) under the percentile restrictions of the previous subsection. However, these rather weak restrictions lead to prior classes that, while plausible, are often too broad for practical use. As such, we might consider a sequence of increasingly tight restrictions on the shape and smoothness of permissible priors, which in turn enable increasingly informative results. For example, we might retain the mixture form (6.7), but now restrict g(θ to some particular parametric family. Carlin and Sargent (1996) refer to such a prior as "semiparametric" since the parametric form for g does not cover the indifference zone $[\theta_L, \theta_U]$, although since we have adopted another parametric form over this range (the uniform) one might argue that "biparametric" or simply "mixture" would be better names.

We leave it as an exercise to show that requiring $G \in \mathcal{H}_c$ is equivalent to requiring $BF \leq \left(\frac{p}{1-p}\right)\left(\frac{1-\pi}{\pi}\right)$, where

$$BF = \frac{\frac{1}{\theta_U - \theta_L}\int_{\theta_L}^{\theta_U} f(x|\theta)d\theta}{\int f(x|\theta)g(\theta)d\theta}\ , \tag{6.8}$$

the Bayes factor in favor of the null hypothesis. Equation (6.8) expresses the Bayes factor as the ratio of the marginal densities under the competing hypotheses; it is also expressible as the ratio of posterior to prior odds in favor of the null. As such, BF gives the extent to which the data have revised our prior beliefs concerning the two hypotheses. Note that if we take $\pi = 1/2$ (equal prior weighting of null and alternative) then a Bayes factor of 1 suggests equal posterior support for the two hypotheses. In this case, we require a Bayes factor of 1/19 or smaller to insure that $P(H_0|x)$ does not exceed 0.05.

In practice, familiar models from the exponential family are often appropriate (either exactly or asymptotically) for the likelihood $f(x|\theta)$. This naturally leads to consideration of the restricted class of *conjugate* priors $g(\theta)$, to obtain a closed form for the integral in the denominator of (6.8). Since a normal approximation to the likelihood for θ is often suitable for even moderate sample sizes, we illustrate in the case of a conjugate normal prior. The fact that g is defined only on the complement of the indifference zone presents a slight complication, but, fortunately, the calculations remain tractable under a renormalized prior with the proper support. That is, we take

$$g(\theta) = \frac{N(\theta|\mu, \tau^2)}{1 - \left[\Phi\left(\frac{\theta_U - \mu}{\tau}\right) - \Phi\left(\frac{\theta_L - \mu}{\tau}\right)\right]}\ ,\quad \theta \notin [\theta_L, \theta_U]\ ,$$

where the numerator denotes the density of a normal distribution with mean μ and variance τ^2, and Φ denotes the cdf of a standard normal distribution.

To obtain a computational form for equation (6.8), suppose we can approximate the likelihood satisfactorily with a $N(\theta|\hat{\theta}, \hat{\sigma}^2)$ density, where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ and $\hat{\sigma}^2$ is a corresponding standard error estimate. Probability calculus then shows that

$$\begin{aligned} \int f(x|\theta)g(\theta)d\theta = & \frac{1}{\sqrt{2\pi(\hat{\sigma}^2+\tau^2)}} \exp\left[-\frac{(\mu-\hat{\theta})^2}{2(\hat{\sigma}^2+\tau^2)}\right] \\ & \times \left\{1 - \left[\Phi\left(\frac{\theta_U-\eta}{\nu}\right) - \Phi\left(\frac{\theta_L-\eta}{\nu}\right)\right]\right\}, \end{aligned} \tag{6.9}$$

where $\eta = (\hat{\sigma}^2\mu + \tau^2\hat{\theta})/(\hat{\sigma}^2 + \tau^2)$ and $\nu^2 = \hat{\sigma}^2\tau^2/(\hat{\sigma}^2 + \tau^2)$. (Note that η and ν^2 are respectively the posterior mean and variance under the fully parametric normal/normal model, described in the next subsection.) Since $\int_{\theta_L}^{\theta_U} f(x|\theta)d\theta = \Phi\left(\frac{\theta_U-\hat{\theta}}{\hat{\sigma}}\right) - \Phi\left(\frac{\theta_L-\hat{\theta}}{\hat{\sigma}}\right)$ we can now obtain the Bayes factor (6.8) without numerical integration, provided that there are subroutines available to evaluate the normal density and cdf.

As a final approach, we might abandon the mixture prior form (6.7) in favor of a single parametric family $h(\theta)$, preferably chosen as conjugate with the likelihood $f(x|\theta)$. If such a choice is possible, we obtain simple closed form expressions for Bayes factors and tail probabilities whose sensitivity to changes in the prior parameters can be easily examined. For example, for our $N(\theta|\hat{\theta}, \hat{\sigma}^2)$ likelihood under a $N(\theta|\mu, \tau^2)$ prior, the posterior probability of H_0 is nothing but

$$P(\theta \in [\theta_L, \theta_U]|x) = \Phi\left(\frac{\theta_U - \eta}{\nu}\right) - \Phi\left(\frac{\theta_L - \eta}{\nu}\right),$$

where η and ν^2 are again as defined beneath equation (6.9). Posterior probabilities that correspond to stopping to reject the hypotheses $H_L : \theta < \theta_L$ and $H_U : \theta > \theta_U$ arise similarly.

6.2 Model assessment

We have already presented several tools for Bayesian model assessment in Subsection 2.4.1. It is not our intention to review all of these tools, but rather to point out how their computation is greatly facilitated by modern Monte Carlo computational methods. Since most Bayesian models encountered in practice require some form of sampling to evaluate the posterior distributions of interest, this means that common model checks will be available at little extra cost, both in terms of programming and runtime.

Consider for example the simple Bayesian residual,

$$r_i = y_i - E(y_i|\mathbf{z}),\ i = 1, \ldots, n,$$

where $\mathbf{z}$ is the sample of data used to fit the model and $\mathbf{y} = (y_1, \ldots, y_n)'$ is an independent validation sample. Clearly, calculation of r_i requires an expectation with respect to the posterior predictive distribution $p(y_i|\mathbf{z})$, which will rarely be available in closed form. However, notice that we can write

$$
\begin{aligned}
E(y_i|\mathbf{z}) &= \int y_i p(y_i|\mathbf{z}) dy_i \\
&= \int\int y_i f(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{z}) d\boldsymbol{\theta} dy_i \\
&= \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{z}) d\boldsymbol{\theta} \\
&\approx \frac{1}{G}\sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)}) \ ,
\end{aligned}
$$

where the $\boldsymbol{\theta}^{(g)}$ are samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{z})$ (for Markov chain Monte Carlo algorithms, we would of course include only post-convergence samples). The equality in the second line holds due to the conditional independence of $\mathbf{y}$ and $\mathbf{z}$ given $\boldsymbol{\theta}$, while that in the third line arises from reversing the order of integration. But $E(y_i|\boldsymbol{\theta})$ typically will be available in closed form, since this is nothing but the mean structure of the likelihood. The fourth line thus arises as a Monte Carlo integration.

Even if $E(y_i|\boldsymbol{\theta})$ is not available in closed form, we can still estimate $E(y_i|\mathbf{z})$ provided we can draw samples $y_i^{(g)}$ from $f(y_i|\boldsymbol{\theta}^{(g)})$. Such sampling is naturally appended onto the algorithm generating the $\boldsymbol{\theta}^{(g)}$ samples themselves. In this case we have

$$
\begin{aligned}
E(y_i|\mathbf{z}) &= \int\int y_i f(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{z}) d\boldsymbol{\theta} dy_i \\
&\approx \frac{1}{G}\sum_{g=1}^{G} y_i^{(g)} \ ,
\end{aligned}
$$

since $(y_i^{(g)}, \boldsymbol{\theta}^{(g)})$ constitute a sample from $p(y_i, \boldsymbol{\theta}|\mathbf{z})$. This estimator will not be as accurate as the first (since both integrals are now being done via Monte Carlo), but will still be simulation consistent (i.e., converge to the true value with probability 1 as $G \to \infty$).

Next, consider the cross-validation residual given in (2.23),

$$
r_i = y_i - E(y_i|\mathbf{y}_{(i)}), \ i = 1, \ldots, n,
$$

where we recall that $\mathbf{y}_{(i)}$ denotes the vector of all the data except the i^{th} value. Now we have

$$
\begin{aligned}
E(y_i|\mathbf{y}_{(i)}) &= \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \\
&\approx \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}
\end{aligned}
$$

$$\approx \frac{1}{G}\sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)}) \ ,$$

where the $\boldsymbol{\theta}^{(g)}$ are samples from the complete data posterior $p(\boldsymbol{\theta}|\mathbf{y})$. The approximation in the second line should be adequate unless the dataset is small and y_i is an extreme outlier. It enables the use of the same $\boldsymbol{\theta}^{(g)}$ samples (already produced by the Monte Carlo algorithm) for estimating each conditional mean $E(y_i|\mathbf{y}_{(i)})$, hence each residual r_i, for $i = 1,\ldots,n$.

Finally, to obtain the cross-validation standardized residual in (2.24), we require not only $E(y_i|\mathbf{y}_{(i)})$, but also $Var(y_i|\mathbf{y}_{(i)})$. This quantity is estimable by first rewriting $Var(y_i|\mathbf{y}_{(i)}) = E(y_i^2|\mathbf{y}_{(i)}) - [E(y_i|\mathbf{y}_{(i)})]^2$, and then observing

$$\begin{aligned} E(y_i^2|\mathbf{y}_{(i)}) &= \int E(y_i^2|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} \\ &= \int \{Var(y_i|\boldsymbol{\theta}) + [E(y_i|\boldsymbol{\theta})]^2\}p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} \\ &\approx \frac{1}{G}\sum_{g=1}^{G}\left\{Var(y_i|\boldsymbol{\theta}^{(g)}) + [E(y_i|\boldsymbol{\theta}^{(g)})]^2\right\} \ , \end{aligned}$$

again approximating the reduced data posterior with the full data posterior and performing a Monte Carlo integration.

Other posterior quantities from Subsection 2.4.1, such as conditional predictive ordinates $f(y_i|\mathbf{y}_{(i)})$ and posterior predictive model checks (e.g., Bayesian p-values) can be derived similarly. The details of some of these calculations are left as exercises.

6.3 Bayes factors via marginal density estimation

As in the previous section, our goal in this section is not to review the methods for model choice and averaging that have already been presented in Subsections 2.3.3 and 2.4.2, respectively. Rather, we provide several computational approaches for obtaining these quantities, especially the Bayes factor. In Section 6.5 we will present some more advanced approaches that are not only implemented via MCMC methods, but also motivated by these methods, in the sense that they are applicable in very highly-parametrized model settings where the traditional model selection method offered by Bayes factors is unavailable or infeasible.

Recall the use of the Bayes factor as a method for choosing between two competing models M_1 and M_2, given in equation (2.20) as

$$BF = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \ , \tag{6.10}$$

the ratio of the observed marginal densities for the two models. For large sample sizes n, the necessary integrals over $\boldsymbol{\theta}_i$ may be available convemently and accurately via asymptotic approximation, producing estimated Bayes factors accurate to order $O(1/n)$, where n is the number of factors contributing to the likelihood function. In any case, equation (6.2) reveals that asymptotic approximations are helpful in discovering the sensitivity of Bayes factors to prior specification, whether they played a role in the original calculation or not.

For moderate sample sizes n or reasonably challenging models, however, such approximations are not appropriate, and sampling-based methods must be used to obtain estimates of the marginal likelihoods needed to evaluate BF. This turns out to be a surprisingly difficult problem: unlike posterior and predictive distributions, marginal distributions are not easily estimated from the output of an MCMC algorithm. As a result, many different approaches have been suggested in the literature, all of which involve augmenting or otherwise "tricking" a sampler into producing the required marginal density estimates. As with our Subsection 5.4.6 discussion of convergence diagnostics, we do not attempt a technical review of every method, but we do comment on their strengths and weaknesses so that the reader can judge which will likely be the most appropriate in a given problem setting. Kass and Raftery (1995) provide a comprehensive review of Bayes factors. including many of the computational methods described below.

6.3.1 Direct methods

In what follows, we suppress the dependence on the model indicator M in our notation, since all our calculations must be repeated for both models appearing in expression (6.10). Observe that since

$$p(\mathbf{y}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \ ,$$

we could generate observations $\{\boldsymbol{\theta}^{(g)}\}_{g=1}^{G}$ from the prior and compute the estimate

$$\hat{p}(\mathbf{y}) = \frac{1}{G} \sum_{g=1}^{G} f(\mathbf{y} \mid \boldsymbol{\theta}^{(g)}) \ , \qquad (6.11)$$

a simple Monte Carlo integration. Unfortunately, the conditional likelihood $f(\mathbf{y} \mid \boldsymbol{\theta})$ will typically be very peaked compared to the prior $p(\boldsymbol{\theta})$, so that (6.11) will be a very inefficient estimator (most of the terms in the sum will be near 0). A better approach would be to use samples from the posterior distribution, as suggested by Newton and Raftery (1994). These authors develop the estimator

$$\hat{p}(\mathbf{y}) = \left[\frac{1}{G} \sum_{g=1}^{G} \frac{1}{f(\mathbf{y} \mid \boldsymbol{\theta}^{(g)})} \right]^{-1} , \qquad (6.12)$$

the harmonic mean of the posterior sample conditional likelihoods. This approach, while efficient, can be quite unstable since a few of the conditional likelihood terms in the sum will still be near 0. Theoretically, this difficulty corresponds to this estimator's failure to obey a Gaussian central limit theorem as $G \rightarrow \infty$. To correct this, Newton and Raftery suggest a compromise between methods (6.11) and (6.12) wherein we use a mixture of the prior and posterior densities for each model, $\tilde{p}(\boldsymbol{\theta}) = \delta p(\boldsymbol{\theta}) + (1-\delta) p(\boldsymbol{\theta}|\mathbf{y})$ as an importance sampling density. Defining $w(\boldsymbol{\theta}) = p(\boldsymbol{\theta})/\tilde{p}(\boldsymbol{\theta})$, we then have

$$\hat{p}(\mathbf{y}) = \frac{\sum_{g=1}^{G} f(\mathbf{y} \mid \boldsymbol{\theta}^{(g)})\, w(\boldsymbol{\theta}^{(g)})}{\sum_{g=1}^{G} w(\boldsymbol{\theta}^{(g)})} .$$

This estimator is more stable and does satisfy a central limit theorem, although in this form it does require sampling from both the posterior and the prior.

A useful generalization of (6.12) was provided by Gelfand and Dey (1994), who began with the identity

$$[p(\mathbf{y})]^{-1} = \int \frac{h(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}\, p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} ,$$

which holds for any proper density h. Given samples $\boldsymbol{\theta}^{(g)}$ from the posterior distribution, this suggests the estimator

$$\hat{p}(\mathbf{y}) = \left[\frac{1}{G} \sum_{g=1}^{G} \frac{h(\boldsymbol{\theta}^{(g)})}{f(\mathbf{y}|\boldsymbol{\theta}^{(g)})\, p(\boldsymbol{\theta}^{(g)})} \right]^{-1} . \tag{6.13}$$

Notice that taking $h(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, the prior density, produces (6.12). However, this is a poor choice from the standpoint of importance sampling theory, which instead suggests choosing h to roughly match the posterior density. Gelfand and Dey suggest a multivariate normal or t density with mean and covariance matrix estimated from the $\boldsymbol{\theta}^{(g)}$ samples. Kass and Raftery (1995) observe that this estimator does satisfy a central limit theorem provided that

$$\int \frac{h^2(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}\, d\boldsymbol{\theta} < \infty ,$$

i.e., the tails of h are relatively thin. The estimator does seem to perform well in practice unless $\boldsymbol{\theta}$ is of too high a dimension. complicating the selection of a good h density.

6.3.2 Using Gibbs sampler output

In the case of models fit via Gibbs sampling with closed-form full conditional distributions, Chib (1995) offers an alternative that avoids the specification of the h function above. The method begins by simply rewriting

Bayes' rule as

$$p(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} \ .$$

Only the denominator on the right-hand side is unknown, so an estimate of the posterior would produce an estimate of the marginal density, as we desire. But since this identity holds for any $\boldsymbol{\theta}$ value, we require only a posterior density estimate at a single point - say. $\boldsymbol{\theta}'$. Thus we have

$$\log p(\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}') + \log p(\boldsymbol{\theta}') - \log p(\boldsymbol{\theta}'|\mathbf{y}) \ , \tag{6.14}$$

where we have switched to the log scale to improve computational accuracy. While in theory $\boldsymbol{\theta}'$ L arbitrary. Chib (1995) suggests choosing it as a point of high posterior density. again to maximize accuracy in (6.14).

It remains to show how to obtain the estimate $\hat{p}(\boldsymbol{\theta}'|\mathbf{y})$. We first describe the technique in the case where the parameter vector can be decomposed into two blocks (similar to the data augmentation scenario of Tanner and Wong. 1987). That is, we suppose that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $p(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1)$ are both available in closed form. Writing

$$p(\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2|\mathbf{y}) = p(\boldsymbol{\theta}'_2|\mathbf{y}, \boldsymbol{\theta}'_1) p(\boldsymbol{\theta}'_1|\mathbf{y}) \ , \tag{6.15}$$

we observe that the first term on the right-hand side is available explicitly at $\boldsymbol{\theta}'$, while the second can be estimated via the "Rao-Blackwellized" mixture estimate (5.16), namely,

$$\hat{p}(\boldsymbol{\theta}'_1|\mathbf{y}) = \frac{1}{G}\sum_{g=1}^{G} p(\boldsymbol{\theta}'_1|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}) \ ,$$

since $\boldsymbol{\theta}_2^{(g)} \sim p(\boldsymbol{\theta}_2|\mathbf{y})$, g = 1, . . ., G. Thus our marginal density estimate in (6.14) becomes

$$\log \hat{p}(\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2) + \log p(\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2) - \log p(\boldsymbol{\theta}'_2|\mathbf{y}, \boldsymbol{\theta}'_1) - \log \hat{p}(\boldsymbol{\theta}'_1|\mathbf{y}) \ .$$

Exponentiating produces the final marginal density estimate.

Next suppose there are three parameter blocks, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$. The decomposition of the joint posterior density (6.15) now becomes

$$p(\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \boldsymbol{\theta}'_3|\mathbf{y}) = p(\boldsymbol{\theta}'_3|\mathbf{y}, \boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2) p(\boldsymbol{\theta}'_2|\mathbf{y}, \boldsymbol{\theta}'_1) p(\boldsymbol{\theta}'_1|\mathbf{y}) \ .$$

Again the first term is available explicitly, while the third term may be estimated as a mixture of $p(\boldsymbol{\theta}'_1|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}, \boldsymbol{\theta}_3^{(g)})$, g = 1, . . ., G, similar to (6.16) above. For the second term, we may write

$$p(\boldsymbol{\theta}'_2|\mathbf{y}, \boldsymbol{\theta}'_1) = \int p(\boldsymbol{\theta}'_2|\mathbf{y}, \boldsymbol{\theta}'_1, \boldsymbol{\theta}_3) p(\boldsymbol{\theta}_3|\mathbf{y}, \boldsymbol{\theta}'_1) d\boldsymbol{\theta}_3 \ ,$$

suggesting the estimator

$$\hat{p}(\boldsymbol{\theta}'_2|\mathbf{y}, \boldsymbol{\theta}'_1) = \frac{1}{G}\sum_{g=1}^{G} p(\boldsymbol{\theta}'_2|\mathbf{y}, \boldsymbol{\theta}'_1, \boldsymbol{\theta}_3^{*(g)}) \ ,$$

where $\boldsymbol{\theta}_3^{*(g)} \sim p(\boldsymbol{\theta}_3|\mathbf{y},\boldsymbol{\theta}_1')$. Such draws are *not* available from the original posterior sample, which instead contains $\boldsymbol{\theta}_3^{(g)} \sim p(\boldsymbol{\theta}_3|\mathbf{y})$. However, we may produce them simply by continuing the Gibbs sampler for an additional G iterations with only two full conditional distributions, namely

$$p(\boldsymbol{\theta}_2|\mathbf{y},\boldsymbol{\theta}_1',\boldsymbol{\theta}_3) \text{ and } p(\boldsymbol{\theta}_3|\mathbf{y},\boldsymbol{\theta}_1',\boldsymbol{\theta}_2)\ .$$

Thus, while additional sampling is required, new computer code is not; we need only continue with a portion of the old code. The final marginal density estimate then arises from

$$\begin{aligned}\log \hat{p}(\mathbf{y}) = &\log f(\mathbf{y}|\boldsymbol{\theta}_1',\boldsymbol{\theta}_2',\boldsymbol{\theta}_3') + \log p(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2',\boldsymbol{\theta}_3') \\ &- \log p(\boldsymbol{\theta}_3'|\mathbf{y},\boldsymbol{\theta}_1',\boldsymbol{\theta}_2') - \log \hat{p}(\boldsymbol{\theta}_2'|\mathbf{y},\boldsymbol{\theta}_1') - \log \hat{p}(\boldsymbol{\theta}_1'|\mathbf{y})\ .\end{aligned}$$

The extension to B parameter blocks requires a similar factoring of the joint posterior into B components, with $(B - 1)$ Gibbs sampling runs of G samples each to estimate the various factors. We note that clever partitioning of the parameter vector into only a few blocks (each still having a closed form full conditional) can increase computational accuracy and reduce programming and sampling time as well.

6.3.3 *Using Metropolis-Hastings output*

Of course, equations like (6.16) require us to know the normalizing constant for the full conditional distribution $p(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2)$, thus precluding their use with full conditionals updated using Metropolis-Hastings (rather than Gibbs) steps. To remedy this, Chib and Jeliazkov (1999) extend the approach, which takes a particularly simple form in the case where the parameter vector $\boldsymbol{\theta}$ can be updated in a single block. Let

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|\mathbf{y}) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\mathbf{y})}{p(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta},\boldsymbol{\theta}^*|\mathbf{y})}\right\}\ ,$$

the probability of accepting a Metropolis-Hastings candidate $\boldsymbol{\theta}^*$ generated from a candidate density $q(\boldsymbol{\theta},\boldsymbol{\theta}^*|\mathbf{y})$ (note that this density is allowed to depend on the data y). Chib and Jeliazkov (1999) then show

$$p(\boldsymbol{\theta}'|\mathbf{y}) = \frac{E_1\left\{\alpha(\boldsymbol{\theta},\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta},\boldsymbol{\theta}'|\mathbf{y})\right\}}{E_2\left\{\alpha(\boldsymbol{\theta}',\boldsymbol{\theta}|\mathbf{y})\right\}}\ , \tag{6.17}$$

where E_1 is the expectation with respect to the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ and E_2 is the expectation with respect to the candidate density q($\boldsymbol{\theta}',\boldsymbol{\theta}|\mathbf{y})$. The numerator is then estimated by averaging the product in braces with respect to draws from the posterior, while the denominator is estimated by averaging the acceptance probability with respect to draws from $q(\boldsymbol{\theta}',\boldsymbol{\theta}|\mathbf{y})$, given the fixed value $\boldsymbol{\theta}'$. Note that this calculation does not require knowledge of the normalizing constant for $p(\boldsymbol{\theta}|\mathbf{y})$. Plugging this estimate of (6.17) into (6.14) completes the estimation of the marginal likelihood. When there are two more more blocks, Chib and Jeliazkov (1999) illustrate an extended version of this algorithm using multiple MCMC runs, similar to the Chib (1995) approach for the Gibbs sampler outlined in the previous subsection.

6.4 Bayes factors via sampling over the model space

The methods of the previous section all seek to estimate the marginal density p(y) for each model, and subsequently calculate the Bayes factor via equation (6.10). They also operate on a posterior sample that has already been produced by some noniterative or Markov chain Monte Carlo method, though the methods of Chib (1995) and Chib and Jeliazkov (1999) will often require multiple runs of slightly- different versions of the MCMC algorithm to produce the necessary- output. But for some complicated or high-dimensional model settings. such as spatial models using Markov random field priors (which involve large numbers of random effect parameters that cannot be analytically integrated out of the likelihood nor readily updated in blocks), these methods may be infeasible.

An alternative approach favored by many- authors is to include the model indicator M as a parameter in the sampling algorithm itself. This of course complicates the initial sampling process. but has the important benefit of producing a stream of samples $\{M^{(g)}\}_{g=1}^{G}$ from $p(M|y)$, the marginal posterior distribution of the model indicator. Hence the ratio

$$\hat{p}(M = j|\mathbf{y}) = \frac{\text{number of } M^{(g)} = j}{\text{total number of } M^{(g)}}, \quad j = 1, \ldots, K \ , \tag{6.18}$$

provides a simple estimate of each posterior model probability, which may then be used to compute the Bayes factor between any two of the models, say j and j', via

$$\widehat{BF}_{jj'} = \frac{\hat{p}(M = j|\mathbf{y})/\hat{p}(M = j'|\mathbf{y})}{p(M = j)/p(M = j')} \ , \tag{6.19}$$

the original formula used to define the Bayes factor in (2.19). Estimated variances of the estimates in (6.18) are easy to obtain even if the $M^{(g)}$ output stream exhibits autocorrelation through the hatching formula (5.39) or other methods mentioned in Subsection 5.4.6.

Sampling orer the model space alone

Most of the methods we will discuss involve algorithmic searches over both the model and the parameter spaces simultaneously. This is of course a natural way to think about the problem, but like any other augmented sampler such an approach risks a less well-identified parameter space, increased correlations, and hence slower convergence than a sampler operating on any one of the models alone. Moreover, if our interest truly lies only in computing posterior model probabilities $p(M = j|\mathbf{y})$ or a Bayes factor, the parameter samples are not needed, relegating the $\boldsymbol{\theta}_j$ to nuisance parameter status.

These thoughts motivate the creation of samplers that operate on the model space alone. This in turn requires us to integrate the $\boldsymbol{\theta}_j$ out of the model before sampling begins. To obtain such marginalized expressions in closed form requires fairly specialized likelihood and prior settings, but several authors have made headway in this area for surprisingly broad classes of models. For example, Madigan and York (1995) offer an algorithm for searching over a space of graphical models for discrete data, an approach they refer to as Markov chain Monte Carlo model composition, or $(MC)^3$. Raftery, Madigan, and Hoeting (1997) work instead in the multiple regression setting with conjugate priors, again enabling a model-space-only search. They compare the $(MC)^3$ approach with the "Occam's Window" method of Madigan and Raftery (1994). Finally, Clyde, DeSimone, and Parmigiani (1996) use importance sampling (not MCMC) to search for the most promising models in a hierarchical normal linear model setting. They employ an orthogonalization of the design matrix which enables impressive gains in efficiency over the Metropolis-based $(MC)^3$ method.

Sampling over model and parameter space

Unfortunately, most model settings are too complicated to allow the entire parameter vector $\boldsymbol{\theta}$ to be integrated out of the joint posterior in closed form, and thus require that any MCMC model search be over the model and parameter space jointly. Such searches date at least to the work of Carlin and Poison (1991), who included M as a parameter in the Gibbs sampler in concert with the scale mixture of normals idea mentioned in Subsection 6.1.1. In this way they computed Bayes factors and compared marginal posterior densities for a parameter of interest under changing specification of the model error densities and related prior densities. Their algorithm required that the models share the same parametrization. however. and so would not be appropriate for comparing two different mean structures (say, a linear and a quadratic). George and McCulloch (1993) circumvented this problem for multiple regression models by introducing latent indicator variables which determine whether or not a particular regression coefficient may be safely estimated by 0. a process they referred to as *stochastic search variable selection* (SSVS). Unfortunately, in order to satisfy the Markov convergence requirement of the Gibbs sampler, a regressor can never completely "disappear" from the model, so the Bayes factors obtained necessarily depend on values of user-chosen tuning constants.

6.4.1 Product space search

Carlin and Chib (1995), whom we abbreviate as "CC", present a Gibbs sampling method that avoids these theoretical convergence difficulties and accommodates completely general model settings. Suppose there are K candidate models, and corresponding to each, there is a distinct parameter vector $\boldsymbol{\theta}_j$ having dimension n_j, $j = 1, \ldots, K$. Our interest lies in $p(M = j \mid \mathbf{y})$, $j = 1, \ldots, K$ the posterior probabilities of each of the K models, and possibly the model-specific posterior distributions $p(\boldsymbol{\theta}_j \mid M = j, \mathbf{y})$ as well.

Supposing that $\boldsymbol{\theta}_j \in \Re^{n_j}$, the approach is essentially to sample over the model indicator and the *Product* space $\prod_{j=1}^{K} \Re^{n_j}$. This entails viewing the prior distributions as model specific and part of the Bayesian model specification. That is, corresponding to model j we write the likelihood as $f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)$ and the prior as $p(\boldsymbol{\theta}_j|M = j)$. Since we are assuming that M merely provides an indicator as to which particular $\boldsymbol{\theta}_j$ is relevant to y, we have that y is independent of $\{\boldsymbol{\theta}_{i\neq j}\}$ given that $M = j$. In addition, since our primary goal is the computation of Bayes factors, we assume that each prior $p(\boldsymbol{\theta}_j|M = j)$ is proper (though possibly quite vague). For simplicity we assume complete independence among the various $\boldsymbol{\theta}_j$ given the model indicator M, and thus may complete the Bayesian model specification by choosing proper "pseudopriors," $p(\boldsymbol{\theta}_j|M \neq j)$. Writing $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, our conditional independence assumptions imply that

$$\begin{aligned} p(\mathbf{y}|M = j) &= \int f(\mathbf{y}|\boldsymbol{\theta}, M = j)p(\boldsymbol{\theta}|M = j)d\boldsymbol{\theta} \\ &= \int f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j)d\boldsymbol{\theta}_j \ , \end{aligned}$$

and so the form given to $p(\boldsymbol{\theta}_j|M \neq j)$ is irrelevant. Thus, as the name suggests, a pseudoprior is not really a prior at all, but only a conveniently chosen linking density, used to completely define the joint model specification. Hence the joint distribution of y and $\boldsymbol{\theta}$ given $M = j$ is

$$p(\mathbf{y}, \boldsymbol{\theta}, M = j) = f(\mathbf{y}|\boldsymbol{\theta}_j, M = j) \left\{ \prod_{i=1}^{K} p(\boldsymbol{\theta}_i|M = j) \right\} \pi_j \ ,$$

where $\pi_j \equiv P(M = j)$ is the prior probability assigned to model j, satisfying $\sum_{j=1}^{K} \pi_j = 1$.

In order to implement the Gibbs sampler, we need the full conditional distributions of each $\boldsymbol{\theta}_j$ and M. The former is given by

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{i\neq j}, M, \mathbf{y}) \propto \begin{cases} f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j) \ , & M = j \\ p(\boldsymbol{\theta}_j|M \neq j) \ , & M \neq j \end{cases} .$$

That is, when $M = j$ we generate from the usual model j full conditional; when $M \neq j$ we generate from the pseudoprior. Both of these generations are straightforward provided $p(\boldsymbol{\theta}_j|M = j)$ is taken to be conjugate with its likelihood. The full conditional for M *is*

$$p(M=j|\boldsymbol{\theta},\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_j, M=j)\left\{\prod_{i=1}^{K} p(\boldsymbol{\theta}_i|M=j)\right\}\pi_j}{\sum_{k=1}^{K} f(\mathbf{y}|\boldsymbol{\theta}_k, M=k)\left\{\prod_{i=1}^{K} p(\boldsymbol{\theta}_i|M=k)\right\}\pi_k} \,. \quad (6.20)$$

Since .11 is a discrete finite parameter, its generation is routine as well. Hence all the required full conditional distributions are well defined and, under the usual regularity conditions (Roberts and Smith, 1993), the algorithm will produce samples from the correct joint posterior distribution. In particular, equations (6.18) and (6.19) can now be used to estimate the posterior model probabilities and the Bayes factor between any two models, respectively.

Notes on implementation

Notice that, in contrast to equation (6.18) summarization of the totality of $\boldsymbol{\theta}_j^{(g)}$ samples is not appropriate. This is because what is of interest in our setting is not the marginal posterior densities $p(\boldsymbol{\theta}_j|\mathbf{y})$, but rather the *conditional* posterior densities $p(\boldsymbol{\theta}_j|M=j,\mathbf{y})$. However, suppose that in addition to $\boldsymbol{\theta}$ we have a vector of nuisance parameters η, common to all models. Then the fully marginal posterior density of η, $p(\eta|\mathbf{y})$, may be of some interest. Since the data are informative about η regardless of the value of M, a pseudoprior for η is not required. Still, care must be taken to insure that η has the same interpretation in both models. For example, suppose we wish to choose between the two nested regression models

$$M=1: \quad y_i = \alpha + \epsilon_i, \qquad \epsilon_i \stackrel{iid}{\sim} \text{Normal}(0,\sigma^2)\,,$$

and

$$M=2: \quad y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \text{Normal}(0,\tau^2)\,,$$

both for $i = 1,\ldots,n$. In the above notation, $\boldsymbol{\theta}_1 = \sigma$, $\boldsymbol{\theta}_2 = (\beta.\tau)$. and $\eta = \alpha$. Notice that α is playing the role of a grand mean in model 1. but is merely an intercept in model 2. The corresponding posteriors could be quite different if, for example, the observed y_i values were centered near 0. while those for x_i were centered far from 0. Besides being unmeaningful from a practical point of view, the resulting $p(\alpha|\mathbf{y})$ would likely be bimodal, a shape that could wreak great havoc with convergence of the MCMC algorithm, since jumps between M = 1 and 2 would be extremely unlikely.

Poor choices of the pseudopriors $p(\boldsymbol{\theta}_j|M \neq j)$ can have a similar deleterious effect on convergence. Good choices will produce $\boldsymbol{\theta}_j^{(g)}$ values that are consistent with the data, so that $p(M=j|\boldsymbol{\theta}.\mathbf{y})$ will still be reasonably large at the next M update step. Failure to generate competitive pseudoprior values will again result in intolerably high autocorrelations in the $M^{(g)}$ chain, hence slow convergence. As such. matching the pseudopriors as nearly as possible to the true model-specific posteriors is recommended.

This can be done using first-order (normal) approximations or other parametric forms designed to mimic the output from K individual preliminary MCMC runs. Note that we are not using the data to help select the prior, but only the pseudoprior.

If for a particular dataset one of the $p(M = j|\mathbf{y})$ is extremely large, the realized chain will exhibit slow convergence due to the resulting near-absorbing state in the algorithm. In this case, the π_j may be adjusted to correct the imbalance; the final Bayes factor computed from (6.10) will still reflect the true odds in favor of $M = j$ suggested by the data.

Finally, one might plausibly consider the modified version of the above algorithm that skips the generation of actual pseudoprior values, and simply keeps $\boldsymbol{\theta}_j^{(g)}$ at its current value when $M^{(g)} \neq j$. While this MCMC algorithm is no longer a Gibbs sampler, Green and O'Hagan (1998) show that it does still converge to the correct stationary distribution (i.e.. its transition kernel does satisfy detailed balance). Under this alternative we would be free to choose $p(\boldsymbol{\theta}_j|M = i) = p(\boldsymbol{\theta}_j|M = j)$ for all i, meaning that all of these terms cancel from the M update step in equation (6.20). This pseudoprior-free version of the algorithm thus avoids the need for preliminary runs, and greatly reduces the generation and storage burdens in the final model choice run. Unfortunately, Green and O'Hagan also find it to perform quite poorly, since without the variability in candidate $\boldsymbol{\theta}_j^{(g)}$ values created by the pseudopriors, switches between models are rare.

6.4.2 "Metropolized" product space search

Dellaportas, Forster, and Ntzoufras (1998) propose a hybrid Gibbs-Metropolis strategy. In their strategy, the model selection step is based on a proposal for a move from model j to j', followed by acceptance or rejection of this proposal. That is, the method is a "Metropolized Carlin and Chib" (MCC) approach, which proceeds as follows;

1. Let the current state be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension n_j.
2. Propose a new model j' with probability $h(j, j')$.
3. Generate $\boldsymbol{\theta}_{j'}$ from a pseudoprior $p(\boldsymbol{\theta}_{j'}|M \neq j')$ as in Carlin and Chib's product space search method.
4. Accept the proposed move (from j to j') with probability

$$\alpha_{j \to j'} = \min\left\{1,\ \frac{f(\mathbf{y}|\boldsymbol{\theta}_{j'}, M = j')p(\boldsymbol{\theta}_{j'}|M = j')p(\boldsymbol{\theta}_j|M = j')\pi_{j'}h(j', j)}{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j)p(\boldsymbol{\theta}_{j'}|M = j)\pi_j h(j, j')}\right\}.$$

Thus by "Metropolizing" the model selection step, the MCC method needs to sample only from the pseudoprior for the proposed model j'. In this method, the move is a Gibbs step or a sequence of Gibbs steps when $j' = j$. Posterior model probabilities and Bayes factors can be estimated as before,

6.4.3 Reversible jump MCMC

This method, originally due to Green (1995), is another strategy that samples over the model and parameter space, but which avoids the full product space search of the Carlin and Chib (1995) method (and the associated pseudoprior specification and sampling), at the cost of a less straightforward algorithm operating on the union space, $\mathcal{M} \times \bigcup_{j \in \mathcal{M}} \Theta_j$. It generates a Markov chain that can "jump" between models with parameter spaces of different dimensions, while retaining the aperiodicity, irreducibility, and detailed balance conditions necessary for MCMC convergence.

A typical reversible jump algorithm proceeds as follows.

1. Let the current state of the Markov chain be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension
2. Propose a new model j' with probability $h(j, j')$.
3. Generate $\mathbf{u}$ from a proposal density $q(\mathbf{u}|\boldsymbol{\theta}_j, j, j')$.
4. Set $(\boldsymbol{\theta}'_{j'}, \mathbf{u}') = \mathbf{g}_{j,j'}(\boldsymbol{\theta}_j, \mathbf{u})$, where $\mathbf{g}_{j,j'}$ is a deterministic function that is 1-1 and onto. This is a "dimension matching" function, specified so that $n_j + dim(\mathbf{u}) = n_{j'} + dim(\mathbf{u}')$.
5. Accept the proposed move (from j to j') with probability $\alpha_{j \to j'}$, which is the minimum of 1 and

$$\frac{f(\mathbf{y}|\boldsymbol{\theta}'_{j'}, M = j')p(\boldsymbol{\theta}'_{j'}|M = j')\pi_{j'}h(j', j)q(\mathbf{u}'|\boldsymbol{\theta}_{j'}, j', j)}{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j)\pi_j h(j, j')q(\mathbf{u}|\boldsymbol{\theta}_j, j, j')} \left| \frac{\partial \mathbf{g}(\boldsymbol{\theta}_j, \mathbf{u})}{\partial(\boldsymbol{\theta}_j, \mathbf{u})} \right| .$$

When $j' = j$, the move can be either a standard Metropolis-Hastings or Gibbs step. Posterior model probabilities and Bayes factors may be estimated from the output of this algorithm as in the previous subsection.

The "dimension matching" aspect of this algorithm (step 4 above) is a bit obscure and merits further discussion. Suppose we are comparing two models, for which $\theta_1 \in \Re$ and $\theta_2 \in \Re^2$. if θ_1 is a subvector of θ_2, then when moving from $j = 1$ to $j' = 2$, we might simply draw $u \sim q(u)$ and set

$$\theta'_2 = (\theta_1, u) .$$

That is, the dimension matching g is the identity function, and so the Jacobian in step 5 is equal to l. Thus if we set $h(1,1) = h(1, 2) = h(2. 1) = h(2,2) = 1/2$, we have

$$\alpha_{1 \to 2} = \min \left\{ 1, \frac{f(\mathbf{y}|\boldsymbol{\theta}'_2, M = 2)p(\boldsymbol{\theta}'_2|M = 2)\pi_2}{f(\mathbf{y}|\boldsymbol{\theta}_1, M = 1)p(\boldsymbol{\theta}_1|M = 1)\pi_1 q(u)} \right\} .$$

with a corresponding expression for $\alpha_{2 \to 1}$.

In many cases, however, θ_1 will not naturally be thought of as a subvector of θ_2. Green (1995) considers the case of a *changepoint model* (see e.g. Table 5.7 and the associated exercises in Chapter 5). in which the choice is between a time series model having a single, constant mean level θ_1, and one having two levels - say, $\theta_{2,1}$ before the changepoint and $\theta_{2,2}$ afterward. In this setting, when moving from Model 2 to 1, we would not likely want to use *either* $\theta_{2,1}$ or $\theta_{2,2}$ as the proposal value θ_1'. A more plausible choice might be

$$\theta_1' = \frac{\theta_{2,1} + \theta_{2,2}}{2} , \tag{6.21}$$

since the average of the pre- and post-changepoint levels should provide a competitive value for the single level in Model 1. To ensure reversibility of this move, when going from Model 1 to 2 we might sample $u \sim q(u)$ and set

$$\theta_{2,1}' = \theta_1 - u \quad \text{and} \quad \theta_{2,2}' = \theta_1 + u ,$$

since this is a 1-1 and onto function corresponding to the deterministic down move (6.21).

Several variations or simplifications of reversible jump MCMC have been proposed for various model classes; see e.g. Richardson and Green (1997) in the context of mixture modeling, and Knorr-Held and Rasser (2000) for a spatial disease mapping application. Also, the "jump diffusion" approach of Phillips and Smith (1996) can be thought of as a variant on the reversible jump idea.

As with other Metropolis-Hastings algorithms, transformations to various parameters are often helpful in specifying proposal densities in reversible jump algorithms (say, taking the log of a variance parameter). It may also be helpful to apply reversible jump to a somewhat reduced model, where we analytically integrate certain parameters out of the model and use (lower-dimensional) proposal densities for the parameters that remain. We will generally not have closed forms for the full conditional distributions of these "leftover" parameters, but this is not an issue since, unlike the Gibbs-based CC method, reversible jump does not require them. Here an example might be hierarchical normal random effects models with conjugate priors: the random effects (and perhaps even the fixed effects) may be integrated out, permitting the algorithm to sample only the model indicator and the few remaining (variance) parameters.

6.4.4 *Using partial analytic structure*

Godsill (1997) proposes use of a "composite model space," which is essentially the setting of Carlin and Chib (1995) except that parameters are allowed to be "shared" between different models. A standard Gibbs sampler applied to this composite model produces the CC method, while a more sophisticated Metropolis-Hastings approach produces a version of the reversible jump algorithm that avoids the "dimension matching" step present in its original formulation (see step 4 in Subsection 6.4.3 above). This step is often helpful for challenging problems (e.g., when moving to a model containing many parameters whose values would not plausibly equal any of those in the current model), but may be unnecessary for simpler problems.

Along these lines, Godsill (1997) outlines a reversible jump method that takes advantage of *partial analytic structure* (PAS) in the Bayesian model. This procedure is applicable when there exists a subvector $(\boldsymbol{\theta}_{j'})_{\mathcal{U}}$ of the parameter vector $\boldsymbol{\theta}_{j'}$ for model j' such that $p((\boldsymbol{\theta}_{j'})_{\mathcal{U}}|(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}, M = j', \mathbf{y})$ is available in closed form, and in the current model j, there exists an equivalent subvector $(\boldsymbol{\theta}_j)_{-\mathcal{U}}$ (the elements of $\boldsymbol{\theta}_j$ *not* in subvector $\mathcal{U}$ of the same dimension as $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}$. Operationally,

1. Let the current state be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension
2. Propose a new model j' with probability $h(j, j')$.
3. Set $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}} = (\boldsymbol{\theta}_j)_{-\mathcal{U}}$.
4. Accept the proposed move with probability

$$\alpha_{j \to j'} = \min\left\{1, \frac{p(j'|(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}, \mathbf{y})h(j', j)}{p(j|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y})h(j, j')}\right\}, \tag{6.22}$$

where $p(j|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y}) = \int p(j, (\boldsymbol{\theta}_j)_{\mathcal{U}}|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y})\ d(\boldsymbol{\theta}_j)_{\mathcal{U}}$.

5. If the model move is accepted update the parameters of the new model $(\boldsymbol{\theta}_{j'})_{\mathcal{U}}$ and $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}$ using standard Gibbs or Metropolis-Hastings steps; otherwise, update the parameters of the old model $(\boldsymbol{\theta}_j)_{\mathcal{U}}$ and $(\boldsymbol{\theta}_j)_{-\mathcal{U}}$ using standard Gibbs or Metropolis-Hastings steps.

Note that model move proposals of the form $j \to j$ always have acceptance probability 1, and therefore when the current model is proposed, this algorithm simplifies to standard Gibbs or Metropolis-Hastings steps. Note that multiple proposal densities may be needed for $(\boldsymbol{\theta}_j)_{\mathcal{U}}$ across models since, while this parameter is common to all of them, its interpretation and posterior support may difFer. Troughton and Godsill (1997) give an example of this algorithm, in which the update step of $(\boldsymbol{\theta}_j)_{\mathcal{U}}$ is skipped when a proposed model move is rejected.

Summary and recommendations

Han and Carlin (2000) review and compare many of the methods described in this and the previous subsections in the context of two examples. the first a simple regression example, and the second a more challenging hierarchical longitudinal model (see Section 7.4). The methods described in this section that sample jointly over model and parameter space (such as product space search and reversible jump) often converge very slowly. due to the difficulty in finding suitable pseudoprior (or proposal) densities. Marginalizing random effects out of the model can be helpful in this regard. but this tends to create new problems: the marginalization often leads to a very complicated form for the model switch acceptance ratio (step 5 in Subsection 6.4.3), thus increasing the chance of an algebraic or computational error. Even with the marginalization, such methods remain difficult to tune. The user will often need a rough idea of the posterior model probabilities $p(M = j|\mathbf{y})$ in order to set the prior model probabilities π_j in such a way that the sampler spends roughly equal time visiting each of the candidate models. Preliminary model-specific runs are also typically required to specify proposal (or pseudoprior) densities for each model.

By contrast, the marginal likelihood methods of Section 6.3 appear relatively easy to program and tune. These methods do not require preliminary runs (only a point of high posterior density, $\boldsymbol{\theta}'$), and in the case of the Gibbs sampler, only a rearrangement of existing computer code. Estimating standard errors is more problematic (the authors' suggested approach involves a spectral density estimate and the delta method), but simply replicating the entire procedure a few times with different random number seeds generally provides an acceptable idea of the procedure's order of accuracy.

In their numerical illustrations, Han and Carlin (2000) found that the RJ and PAS methods ran more quickly than the other model space search methods, but the marginal likelihood methods seemed to produce the highest degree of accuracy for roughly comparable runtimes. This is in keeping with the intuition that some gain in precision should accrue to MCMC methods that avoid a model space search.

As such, we recommend the marginal likelihood methods as relatively easy and safe approaches when choosing among a collection of standard (e.g., hierarchical linear) models. We hasten to add however that the blocking required by these methods may preclude their use in some settings, such as spatial models using Markov random field priors (which involve large numbers of random effect parameters that cannot be analytically integrated out of the likelihood nor readily updated in blocks; see Subsection 7.8.2). In such cases, reversible jump may offer the only feasible alternative for estimating a Bayes factor. The marginal likelihood methods would also seem impractical if the number of candidate models were very large (e.g., in variable selection problems having 2^p possible models, corresponding to each of p predictors being either included or excluded). But as alluded to earlier, we caution that the ability of joint model and parameter space search methods to sample effectively over such large spaces is very much in doubt; see for example Clyde et al. (1996).

6.5 Other model selection methods

The Bayes factor estimation methods discussed in the previous two sections require substantial time and effort (both human and computer) for a rather modest payoff, namely, a collection of posterior model probability estimates (possibly augmented with associated standard error estimates). Besides being mere single number summaries of relative model worth, Bayes factors are not interpretable with improper priors on any components of the parameter vector, since if the prior distribution is improper then the marginal distribution of the data necessarily is as well. Even for proper priors, the Bayes factor has been criticized on theoretical grounds; see for example Gelfand and Dey (1994) and Draper (1995). One might conclude that none of the methods considered thus far is appropriate for everyday, "rough and ready" model comparison, and instead search for more computationally realistic alternatives.

One such alternative might be more informal, perhaps graphical methods for model selection. These could be based on the marginal distributions $m(y_i)$, $i = 1,...,n$ (as in Berger 1985. p. 199). provided they exist. Alternatively, we could use the conditional predictive distributions $f(y_i|\mathbf{y}_{(i)})$, $i = 1, \ldots, n$ (as in Gelfand et al., 1992), since they will be proper when $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ is, and they indicate the likelihood of each datapoint in the presence of all the rest. For example, the product (or the sum of the logs) of the observed conditional predictive ordinate (CPO) values $f(y_i^{obs}|\mathbf{y}_{(i)})$ given in (2.25) could be compared across models. with the larger result indicating the preferred model. Alternatively, sums of the squares or absolute values of the standardized residuals d_i given in (2.24) could be compared across models, with the model having the smaller value now being preferred. We could improve this procedure by accounting for differing model size - say, by plotting the numerator of (2.24) versus its denominator for a variety of candidate models on the same set of axes. In this way we could judge the accuracy of each model relative to its precision. For example, an overfitted model (i.e., one with redundant predictor variables) would tend to have residuals of roughly the same size as an adequate one. but with higher variances (due to the "collinearity" of the predictors). We exemplify this conditional predictive approach in our Section 8.1 case study.

6.5.1 Penalized likelihood criteria

For some advanced models, even cross-validatory predictive selection methods may be unavailable. For example, in our Section 8.3 case study, the presence of certain model parameters identified only by the prior leads to an information deficit that causes the conditional predictive distributions (2.25) to be improper. In many cases, informal likelihood or penalized likelihood criteria may offer a feasible alternative. Loglikelihood summaries are easy to estimate using posterior samples $\{\boldsymbol{\theta}^{(g)}, g = 1.....G\}$. since we may think of $\ell \equiv \log L(\boldsymbol{\theta})$ as a parametric function of interest. and subsequently compute

$$\hat{\ell} \equiv E[\log L(\boldsymbol{\theta})|\mathbf{y}] \approx \frac{1}{G}\sum_{g=1}^{G} \log L(\boldsymbol{\theta}^{(g)})$$

as an overall measure of model fit to be compared across models. To account for differing model size, we could penalize $\hat{\ell}$ using the same sort of penalties as the Bayesian information (Schwarz) criterion (2.21) or the Akaike information criterion (2.22). In the case of the former, for example, we would have

$$\widehat{BIC} = 2\hat{\ell} - p \log n \ ,$$

where as usual p is the number of parameters in the model, and n is the number of datapoints.

Unfortunately, a problem arises here for the case of hierarchical models: what exactly are p and n? For example, in a longitudinal setting (see Section 7.4) in which we have s_i observations on patient i, $i = 1, \ldots, m$, shall we set $n = \sum_i s_i$ (the total number of observations), or $n = m$ (the number of patients)? If the observations on every patient were independent, the former choice would seem most appropriate, while if they were perfectly correlated within each patient, we might instead choose the latter. But of course the true state of nature is likely somewhere in between. Similarly, if we had a collection of rn random effects, one for each patient, what does this contribute to p? If the random effects had nothing in common (i.e., they were essentially like fixed effects), they would contribute the full m parameters to p, but if the data (or prior) indicated they were all essentially identical, they would contribute little more than one "effective parameter" to the total model size p. Pauler (1998) obtains results in the case of hierarchical normal linear models, but results for general models remain elusive.

In part to help address this problem, Spiegelhalter et al. (1998) suggest a generalization of the Akaike information criterion (AIC) that is based on the posterior distribution of the *deviance* statistic,

$$D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y}|\boldsymbol{\theta}) + 2 \log h(\mathbf{y}) \ , \tag{6.23}$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function for the observed data vector $\mathbf{y}$ given the parameter vector $\boldsymbol{\theta}$, and $h(\mathbf{y})$ is some standardizing function of the data alone (which thus has no impact on model selection). In this approach, the fit of a model is summarized by the posterior expectation of the deviance, $\overline{D} = E_{\theta|y}[D]$, while the *complexity* of a model is captured by the effective number of parameters p_D (note that this is often less than the total number of model parameters due to the borrowing of strength across individual-level parameters in hierarchical models). Below we show that a reasonable definition of p_D is the expected deviance minus the deviance evaluated at the posterior expectations,

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\boldsymbol{\theta}]) = \overline{D} - D(\bar{\boldsymbol{\theta}}) \ . \tag{6.24}$$

The *Deviance Information Criterion* (DIC) is then defined as

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\bar{\boldsymbol{\theta}}) \ , \tag{6.25}$$

with smaller values of DIC indicating a better-fitting model. As with other penalized likelihood criteria, we caution that DIC is not intended for identification of the "correct" model, but rather merely as a method of comparing a collection of alternative formulations (all of which may be incorrect).

An asymptotic justification of DIC is straightforward in cases where the number of observations n grows with respect to the number of parameters p, and where the prior $p(\theta)$ is non-hierarchical and completely specified (i.e., having no unknown parameters). Here we may expand $D(\boldsymbol{\theta})$ around $\bar{\boldsymbol{\theta}}$ to give, to second order,

$$D(\boldsymbol{\theta}) \approx D(\bar{\boldsymbol{\theta}}) - 2(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T L' - (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T L''(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \qquad (6.26)$$

where $L = \log p(\mathbf{y}|\boldsymbol{\theta}) = -D(\boldsymbol{\theta})/2$, and L' and L'' are the first derivative vector and second derivative matrix with respect to $\boldsymbol{\theta}$. But from the Bayesian Central Limit Theorem (Theorem 5.1) we have that $\boldsymbol{\theta}|\mathbf{y}$ is approximately distributed as $N\left(\hat{\boldsymbol{\theta}}, [-L'']^{-1}\right)$, where $\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ are the maximum likelihood estimates such that $L' = 0$. This in turn implies that $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T(-L'')(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ has an approximate chi-squared distribution with p degrees of freedom. Thus writing $D_{\mathrm{non}}(\boldsymbol{\theta})$ to represent the deviance for a non-hierarchical model, from (6.26) we have that

$$D_{\mathrm{non}}(\boldsymbol{\theta}) \approx D(\hat{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T L''(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = D(\hat{\boldsymbol{\theta}}) + \chi^2_p \, .$$

Rearranging this expression and taking expectations with respect to the posterior distribution of $\boldsymbol{\theta}$, we have

$$p \approx E_{\theta|y}[D_{\mathrm{non}}(\boldsymbol{\theta})] - D(\hat{\boldsymbol{\theta}}) \, , \qquad (6.27)$$

so that the number of parameters is approximately the expected deviance $\overline{D} = E_{\theta|y}[D_{\mathrm{non}}(\boldsymbol{\theta})]$ minus the fitted deviance. But since AIC = $D(\boldsymbol{\theta}) + 2p$. from (6.27) we obtain $AIC \approx \overline{D} + p$, the expected deviance plus the number of parameters. The DIC approach for hierarchical models thus follows this equation and equation (6.27), but substituting the posterior mean $\bar{\theta}$ for the maximum likelihood estimate $\hat{\theta}$. It is a generalisation of Akaike's criterion. since for non-hierarchical models $\bar{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$, $p_D \approx p$, and DIC $\approx$ AIC. DIC can also be shown to have much in common with the hierarchical model selection tools previously suggested by Ye (1998) and Hodges and Sargent (1998), though the DIC idea applies much more generally.

As with all penalized likelihood criteria, DIC consists of two terms, one representing "goodness of fit" and the other a penalty for increasing model complexity. DIC is also scale-free; the choice of standardizing function $h(\mathbf{y})$ in (6.23) is arbitrary. Thus values of DIC have no intrinsic meaning; as with AIC and BIC, only *differences* in DIC across models are meaningful. (Of course, p_D *does* have a scale, namely, the size of the effective parameter space.) Besides its generality, another attractive aspect of DIC is that it may be readily calculated during an MCMC run by monitoring both θ and $D(\boldsymbol{\theta})$, and at the end of the run simply taking the sample mean of the simulated values of D, minus the plug-in estimate of the deviance using the sample means of the simulated values of $\boldsymbol{\theta}$. This quantity can be calculated for each model being considered without analytic adaptation, complicated loss functions, additional MCMC sampling (say, of predictive values), or any matrix inversion.

While DIC has already been profitably applied in practice (see e.g. Erkanli et al., 1999), as with any new data analysis technique, several practical questions must be resolved before DIC can become a standard element of the applied Bayesian toolkit. For example, DIC is not invariant to parametrization, so (as with prior elicitation) the most plausible parametrization must be carefully chosen beforehand. Unknown scale parameters and other innocuous restructuring of the model can also lead to small changes in the computed DIC value. Finally, determining an appropriate variance estimate for DIC remains a vexing practical problem. Zhu and Carlin (2000) experiment with various delta method approaches to this problem in the context of spatio-temporal models of the sort considered in Section 7.8.2, but conclude that a "brute force" replication approach may be the only suitably accurate method in such complicated settings. That is, we would independently replicate the calculation of DIC a large number of times N, obtaining a sequence of DIC estimates $\{DIC_l,\ l = 1,\ldots,N\}$, and estimate $Var(DIC)$ by the sample variance,

$$\widehat{Var}(DIC) = \frac{1}{N-1}\sum_{l=1}^{N}(DIC_l - \overline{DIC})^2 .$$

6.5.2 *Predictive model selection*

Besides being rather ad hoc, a problem with penalized likelihood approaches is that the usual choices of penalty- function are motivated by asymptotic arguments. But for many of the complex and high-dimensional models we currently envision, the dimension of the parameter space p increases with the sample size n, so the usual asymptotic calculations are inappropriate. While DIC does offer a way to estimate the effective sizes of such non-regular models, we might instead work purely in predictive space, following Laud and Ibrahim (1995). Here we shall see that intuitively appealing penalties emerge without the need for asymptotics. As in Chapter 2, the basic distribution we work with is

$$f(\mathbf{y}_{new}|\mathbf{y}_{obs}) = \int f(\mathbf{y}_{new}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y}_{obs})\, d\boldsymbol{\theta} , \qquad (6.28)$$

where $\boldsymbol{\theta}$ denotes the collection of model parameters, and $\mathbf{y}_{new}$ is viewed as a replicate of the observed data vector $\mathbf{y}_{obs}$. The model selection criterion first selects a discrepancy function $d(\mathbf{y}_{new}, \mathbf{y}_{obs})$, then computes

$$E[d(\mathbf{y}_{new}, \mathbf{y}_{obs})|\mathbf{y}_{obs}, M_i] \ , \tag{6.29}$$

and selects the model which minimizes (6.29). For Gaussian likelihoods, Laud and Ibrahim (1995, p. 250) suggested

$$d(\mathbf{y}_{new}, \mathbf{y}_{obs}) = (\mathbf{y}_{new} - \mathbf{y}_{obs})^T(\mathbf{y}_{new} - \mathbf{y}_{obs}) \ . \tag{6.30}$$

For a non-Gaussian generalized linear mixed model, we may prefer to replace (6.30) by the corresponding deviance criterion. For example, with a Poisson likelihood we would set

$$d(\mathbf{y}_{new}, \mathbf{y}_{obs}) = 2\sum_l \{y_{l,obs} \log(y_{l,obs}/y_{l,new}) - (y_{l,obs} - y_{l,new})\} \ , \tag{6.31}$$

where l indexes the components of $\mathbf{y}$. Routine calculation shows that the l^{th} term in the summation in (6.31) is strictly convex in $y_{l,new}$ if $y_{l,obs} > 0$. To avoid problems with extreme values in the sample space, we replace (6.31) by

$$\begin{aligned} &\tilde{d}(\mathbf{y}_{new}, \mathbf{y}_{obs}) \\ &\quad = 2\textstyle\sum_l \left\{ (y_{l,obs} + \frac{1}{2}) \log\left(\frac{y_{l,obs}+\frac{1}{2}}{y_{l,new}+\frac{1}{2}}\right) - (y_{l,obs} - y_{l,new}) \right\} \ . \end{aligned}$$

Suppose we write

$$\begin{aligned} E[\tilde{d}(\mathbf{y}_{new}, \mathbf{y}_{obs}) \mid \mathbf{y}_{obs}, M_i] \ &= \ \tilde{d}(E[\mathbf{y}_{new}|\mathbf{y}_{obs}, M_i], \mathbf{y}_{obs}) \\ &+E[\tilde{d}(\mathbf{y}_{new}, \mathbf{y}_{obs}) \mid \mathbf{y}_{obs}, M_i] - \tilde{d}(E[\mathbf{y}_{new}|\mathbf{y}_{obs}, M_i], \mathbf{y}_{obs}) \ . \end{aligned} \tag{6.32}$$

Intuitive interpretations may be given to the terms in (6.32). The left-hand side is the expected predictive deviance (EPD) for model M_i. The first term on the right-hand side is essentially the likelihood ratio statistic with the MLE for $E(\mathbf{y}_{new}|\boldsymbol{\theta}_i, M_i)$, which need not exist, replaced by $E(\mathbf{y}_{new}|\mathbf{y}_{obs}, M_i)$. Jensen's inequality shows that the second term minus the third is strictly positive, and is the penalty associated with M_i. This difference becomes

$$\begin{aligned} &2\textstyle\sum_l \left(y_{l,obs} + \frac{1}{2}\right) \\ &\quad \times \left\{\log E\left[y_{l,new} + \tfrac{1}{2} \middle| \mathbf{y}_{obs}\right] - E\left[\log\left(y_{l,new} + \tfrac{1}{2}\right) \middle| \mathbf{y}_{obs}\right]\right\} \ . \end{aligned} \tag{6.33}$$

Again, each term in the summation is positive. A second order Taylor series expansion shows that (6.33) is approximately

$$\sum_l \frac{y_{l,obs} + \frac{1}{2}}{[E(y_{l,new} + \frac{1}{2} \mid \mathbf{y}_{obs})]^2} \cdot Var(y_{l,new} \mid \mathbf{y}_{obs}) \ . \tag{6.34}$$

Hence (6.33) can be viewed as a weighted predictive variability penalty, a natural choice in that if M_i is too large (i.e., contains too many explanatory terms resulting in substantial multicollinearity), predictive variances will increase. Lastly, (6.34) is approximately

$$E\left\{\sum_l \frac{(y_{l,new} - E(y_{l,new}|\mathbf{y}_{obs}))^2}{E(y_{l,new} + \frac{1}{2} \mid \mathbf{y}_{obs})}\middle|\mathbf{y}_{obs}\right\}$$

and thus may be viewed as a predictive *corrected* goodness-of-fit statistic.

Computation of (6.32) requires calculation of $E(y_{l,new}|\mathbf{y}_{obs}, M_i)$ as well as $E[\log(y_{l,new} + \frac{1}{2})|\mathbf{y}_{obs}, M_i]$. Such predictive expectations are routinely obtained as Monte Carlo integrations; the case study in Section 8.3 provides an example.

Finally, we remark that Gelfand and Ghosh (1998) extend the above approach to a completely formal decision-theoretic setting. Their method seeks to choose the model minimizing a particular posterior predictive loss. Like the less formal method above, the resulting criterion consists of a (possibly weighted) sum of a goodness-of-fit term and a model complexity penalty term.

6.6 Exercises

1. Consider the data displayed in Table 6.1, originally collected by Treloar (1974) and reproduced in Bates and Watts (1988). These data record the "velocity" y_i of an enzymatic reaction (in counts/min/min) as a function of substrate concentration x_i (in ppm), where the enzyme has been treated with Puromycin.

Case (i)	x_i	y_i	Case (i)	x_i	y_i
1	0.02	76	7	0.22	159
2	0.02	47	8	0.22	152
3	0.06	97	9	0.56	191
4	0.06	107	10	0.56	201
5	0.11	123	11	1.10	207
6	0.11	139	12	1.10	200

Table 6.1 *Puromycin experiment data*

A common model for analyzing biochemical kinetics data of this type is the *Michaelis-Menten* model, wherein we adopt the mean structure

$$\mu_i = \gamma + \alpha x_i/(\theta + x_i) ,$$

where $\alpha, \gamma \in \Re$ and $\theta \in \Re^+$. In the nomenclature of Example 6.2, we have design matrix

$$F_\theta^T = \begin{pmatrix} 1 & \cdots & 1 \\ \frac{x_1}{\theta+x_1} & \cdots & \frac{x_n}{\theta+x_n} \end{pmatrix}$$

and $\boldsymbol{\beta}^T = (\gamma, \alpha)$. For this model, obtain estimates of the marginal posterior density of the parameter α, and also the marginal posterior density of the mean velocity at $X = 0.5$, a substrate concentration not represented in the original data, assuming that the error density of the data is

(a) normal

(b) Student's t with 2 degrees of freedom

(c) double exponential.

2. For the point null prior partitioning setting described in Subsection 6.1.2, show that requiring $G \in \mathcal{H}_c$ as given in (6.4) is equivalent to requiring $BF \leq \left(\frac{p}{1-p}\right)\left(\frac{1-\pi}{\pi}\right)$, where BF is the Bayes factor in favor of the null hypothesis.

3. For the interval null prior partitioning setting of Subsection 6.1.2, derive an expression for the set of priors $\mathcal{H}_c$ that correspond to rejecting H_0, similar to expression (6.4) for the point null case.

4. Suppose we are estimating a set of residuals r_i using a Monte Carlo approach, as described in Section 6.2. Due to a small sample size n, we are concerned that the approximation

$$E(y_i|\mathbf{y}_{(i)}) = \int E(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} \approx \int E(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

will not be accurate. A faculty member suggests the importance sampling estimate

$$\begin{aligned} E(y_i|\mathbf{y}_{(i)}) &= \int E(y_i|\boldsymbol{\theta})\frac{p(\boldsymbol{\theta}|\mathbf{y}_{(i)})}{p(\boldsymbol{\theta}|\mathbf{y})}p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &\approx \frac{1}{G}\sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)})\frac{p(\boldsymbol{\theta}^{(g)}|\mathbf{y}_{(i)})}{p(\boldsymbol{\theta}^{(g)}|\mathbf{y})} \end{aligned}$$

as an alternative. Is this a practical solution? What other approaches might we try? (*Hint:* See equation (5.8).)

5. Suppose we have a convergent MCMC algorithm for drawing samples from $p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. We wish to locate potential outliers by computing the conditional predictive ordinate $f(y_i|\mathbf{y}_{(i)})$ given in equation (2.25) for each $i = 1, \ldots, n$. Give a computational formula we could use to obtain a simulation-consistent estimate of $f(y_i|\mathbf{y}_{(i)})$. What criterion might we use to classify y_i as a suspected outlier?

6. In the previous problem, suppose we wish to evaluate the model using the model check p'_D given in equation (2.28) with an independent validation data sample **z**. Give a computational formula we could use to obtain a simulation-consistent estimate of p'_D.

7. Consider again the data in Table 5.4. Define model 1 to be the three variable model given in Example 5.7, and model 2 to be the reduced model having $m_1 = 1$ (i.e., the standard logistic regression model). Compute the Bayes factor in favor of model 1 using

   (a) the harmonic mean estimator, (6.12)

   (b) the importance sampling estimator, (6.13).

   Does the second perform better, as expected?

8. Consider the dataset of Williams (1959), displayed in Table 6.2. For $n = 42$ specimens of radiata pine, the maximum compressive strength parallel to the grain y_i was measured, along with the specimen's density, x_i, and its density adjusted for resin content, z_i (resin contributes much to density but little to strength of the wood). For $i = 1,\ldots,n$ we wish to compare the two models $M = 1$ and $M = 2$ where

$$M = 1: \quad y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)\ ,$$

and

$$M = 2: \quad y_i = \gamma + \delta(z_i - \bar{z}) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \tau^2)\ .$$

We desire prior distributions that are roughly centered around the appropriate least squares parameter estimate, but are extremely vague (though still proper). As such, we place

$$N\left((3000, 185)^T,\ Diag(10^6, 10^4)\right)$$

priors on $(\alpha, \beta)^T$ and $(\gamma, \delta)^T$ and

$$IG\left(3\ ,\ (2 \cdot 300^2)^{-1}\right)$$

priors on σ^2 and τ^2 having both mean and standard deviation equal to 300^2.

Compute the Bayes factor in favor of model 2 (the adjusted density model) using

   (a) the marginal density estimation approach of Subsection 6.3.2

   (b) the product space search approach of Subsection 6.4.1

   (c) the reversible jump approach of Subsection 6.4.3.

How do the methods compare in terms of accuracy? Ease of use? Which would you attempt first in a future problem?

Case (i)	y_i	x_i	z_i	Case (i)	y_i	x_i	z_i
1	3040	29.2	25.4	22	3840	30.7	30.7
2	2470	24.7	22.2	23	3800	32.7	32.6
3	3610	32.3	32.2	24	4600	32.6	32.5
4	3480	31.3	31.0	25	1900	22.1	20.8
5	3810	31.5	30.9	26	2530	25.3	23.1
6	2330	24.5	23.9	27	2920	30.8	29.8
7	1800	19.9	19.2	28	4990	38.9	38.1
8	3110	27.3	27.2	29	1670	22.1	21.3
9	3160	27.1	26.3	30	3310	29.2	28.5
10	2310	24.0	23.9	31	3450	30.1	29.2
11	4360	33.8	33.2	32	3600	31.4	31.4
12	1880	21.5	21.0	33	2850	26.7	25.9
13	3670	32.2	29.0	34	1590	22.1	21.4
14	1740	22.5	22.0	35	3770	30.3	29.8
15	2250	27.5	23.8	36	3850	32.0	30.6
16	2650	25.6	25.3	37	2480	23.2	22.6
17	4970	34.5	34.2	38	3570	30.3	30.3
18	2620	26.2	25.7	39	2620	29.9	23.8
19	2900	26.7	26.4	40	1890	20.8	18.4
20	1670	21.1	20.0	41	3030	33.2	29.4
21	2540	24.1	23.9	42	3030	28.2	28.2

Table 6.2 *Radiata pine compressive strength data*

CHAPTER 7

Special methods and models

Having learned the basics of specifying, fitting, evaluating, and comparing Bayes and EB models, we are now well equipped to tackle real-world data analysis. However, before doing this (in Chapter 8), we outline the Bayesian treatment of several special methods and models useful in practice. We make no claim of comprehensive coverage; indeed, each of this chapter's sections could be expanded into a book of its own! Rather, our goal is to highlight the key differences between the Bayes/EB and classical approaches in these important areas, and provide links to more complete investigations. As general Bayesian modeling references, we list the recent books by O'Hagan (1994) and Gelman, Carlin, Stern, and Rubin (1995).

7.1 Estimating histograms and ranks

Much of this book has concentrated on inferences for coordinate-specific parameters: an experimental unit, an individual, a clinic, a geographic region, etc. We now consider situations where the collection of parameters (the *ensemble*) is of primary or equal interest. For example, policy-relevant environmental or public health assessments involve synthesis of information (e.g. disease rates, relative risks, operative death rates) from a set of related geographic regions. Most commonly, region-specific rates are estimated and used for a variety of assessments. However, estimates of the histogram or *empirical distribution function* (edf) of the underlying rates (for example, to evaluate the number of rates above a threshold) or their rank ordering (for example, to prioritize environmental assessments) can be of equal importance. Other settings in which histograms or ranks might be of interest include estimating clinic-specific treatment effects, hospital or physician specific surgical death rates, school-specific teaching effectiveness, or county-specific numbers of children in poverty. For consistency, throughout this section we will always refer to the units as "regions."

Broadly put, our goal is to estimate the edf of the latent parameters that generated the observed data. One may also be interested in the individual parameters, but the initial goal is estimation of the edf (see Tukey, 1974; Louis, 1984; Laird and Louis, 1989; and Devine, Halloran, Louis, 1994). As shown by Shen and Louis (1998), Conlon and Louis (1999), and references therein, while posterior means are the most obvious and effective estimates for region-specific parameters, their edf is underdispersed and never valid for estimating the edf (or histogram) of the parameters, while the edf of the coordinate-specific MLEs is overdispersed. Ranking the posterior means can also produce suboptimal estimates of the parameter ranks; effective estimates of the parameter ranks should target them directly. Fortunately, structuring via Bayesian decision theory guides development of effective estimates of histograms and ranks, as we now describe.

7.1.1 *Model and inferential goals*

Consider again the two-stage, hierarchical model with a continuous prior distribution:

$$\begin{aligned} \theta_i &\overset{iid}{\sim} G, \\ Y_i \mid \theta_i &\overset{ind}{\sim} f_i(Y_i \mid \theta_i), \end{aligned} \tag{7.1}$$

for $j = 1, \ldots, k$. Let g be the density function of G. Although other loss functions may be more appropriate, to structure estimates in a mathematically convenient manner we use squared error loss (SEL) for the relevant estimation goal. Our goals are to estimate the individual θ_i, their ranks, and the histogram associated with their edf; we now consider each in turn.

Estimating the Individual θ*s*

Under squared error loss, the posterior means (PM),

$$\hat{\theta}_i^{PM} \equiv \eta_i = E[\theta_i \mid Y_i] \,.$$

are optimal. These are the "traditional" parameter estimates. In our subsequent development, all estimates are posterior means of target features, but we reserve the "PM" label for estimates of the individual θs.

Ranking

Laird and Louis (1989) and Stern and Cressie (1999) study the estimation of ranks. The rank vector $\mathbf{R} = (R_1, \ldots, R_k)$ has components $R_i = \text{rank}(\theta_i) = \sum_{l=1}^{k} I_{\{\theta_i \geq \theta_l\}}$ where $I_{\{\cdot\}}$ is the indicator function. The smallest θ has rank 1. The posterior mean rank, $\bar{R}_i$, is optimal under squared error loss:

$$\bar{R}_i = E[R_i \mid \mathbf{Y}] = \sum_{\ell=1}^{k} P[\theta_i \geq \theta_\ell \mid \mathbf{Y}] \,.$$

See Laird and Louis (1989) for the posterior covariance of the ranks.

Generally, the $\bar{R}_i$ are not integers. Though their mean, $(k+1)/2$, is the same as for integer ranks, their spread is smaller, and often substantially smaller. For example, the largest $\bar{R}_i$ can be far smaller than k and the smallest can be far larger than 1. The non-integer feature of the $\bar{R}_i$ is attractive, because integer ranks can overrepresent distance and underrepresent uncertainty, but we will need integer ranks. To obtain them, rank the $\bar{R}_i$, producing

$$\hat{R}_i = \text{rank}(\bar{R}_i) \ .$$

Estimating. the EDF

Let G_k denote the edf of the θs that underlie the current data set. Then $G_k(t \mid \boldsymbol{\theta}) = \frac{1}{k}\sum_{i=1}^{k} I_{\{\theta_i \leq t\}}, -\infty < t < \infty$. With $A(t)$ a candidate estimate of $G_k(t)$, under weighted integrated squared error loss, $\int w(t)[A(t) - G_k(t)]^2 dt$, the posterior mean is optimal:

$$\bar{G}_k(t) \quad = \quad E[G_k(t) \mid \mathbf{Y}] = \frac{1}{k}\sum P(\theta_i \leq t \mid Y_i).$$

With the added constraint that $A(t)$ is a discrete distribution with at most k mass points, Shen and Louis (1998) show that the optimal estimator $\hat{G}_k$ has mass $\frac{1}{k}$ at

$$\hat{U}_i = \bar{G}_k^{-1}\left(\frac{2j-1}{2k}\right), \ j = 1, \ldots, k. \tag{7.2}$$

We now show that both the PMs and the Ys themselves are inappropriate as estimates of G_k. Except when θ is the conditional mean of Y_i, the cdf based on the Ys may have no direct relevance to the distribution of θ. Even when the Ys are relevant (when θ is the mean parameter), the histogram produced from the Y_i is overdispersed. It has the appropriate mean, but its variance is $E[V(Y \mid \theta)] + V[E(Y \mid \theta)] = E[V(Y \mid \theta)] + V[\theta]$; the first term produces overdispersion.

On the other hand, the histogram produced from the PMs is underdispersed. To see this, let $v_i = V(\theta_i \mid Y_i)$ and recall $\eta_i = E(\theta \mid Y_i)$. Then the moments produced by $\bar{G}_k$ (and approximately those produced by $\hat{G}_k$) are

$$\begin{aligned} \text{mean} \quad &= \quad \int t d\bar{G}_k(t) = \frac{1}{k}\sum \eta_i = \bar{\eta}, \\ \text{variance} \quad &= \quad \int (t - \bar{\eta})^2 \, d\bar{G}_k(t) \\ & \quad \frac{1}{k}\sum v_i + \frac{1}{k}\sum (\eta_i - \bar{\eta})^2. \end{aligned} \tag{7.3}$$

The sample variance of the PMs produces only the second term, so the histogram based on them is underdispersed. To correct for this underdispersion, Louis (1984) and Ghosh (1992) derived *constrained Bayes* (CB) estimates. These estimates minimize posterior expected (weighted) SEL for individual parameters subject to the constraint that their sample mean and variance match those in (7.3). Specifically, CB estimates $(a_1, \ldots, a_k)$

minimize

$$\sum w_i E[(a_i - \theta_i)^2 \mid \mathbf{Y}] = \sum w_i(a_i - \eta_i)^2 \ + \sum w_i v_i,$$

subject to the constraints

$$\begin{aligned} \frac{1}{k}\sum a_i &= \frac{1}{k}\sum \eta_i \\ \frac{1}{k}\sum (a_i - \bar{a})^2 &= \frac{1}{k}\sum v_i \ + \ \frac{1}{k}\sum(\eta_i - \bar{\eta})^2. \end{aligned}$$

The equally-weighted case $(w_i \equiv 1)$ produces the closed-form solution

$$\hat{\theta}_i^{CB} = \bar{\eta} + \left[1 + \frac{\sum v_i}{\sum(\eta_i - \bar{\eta})^2}\right]^{\frac{1}{2}} (\eta_i - \bar{\eta}).$$

These CB estimates start with the PMs and adjust them to have the variance in (7.3). Since the term in square brackets is greater than l. the estimates are more spread out around $\bar{\eta}$ than are the PMs. Table 3.3 and Figure 3.1 present an example of these estimates.

Building on the CB approach, Shen and Louis (1998) defined a "G then rank" (GR) approach by assigning the mass points of $\hat{G}_k$ (equation (7.2) to coordinates using the ranks $\hat{R}$,

$$\hat{\theta}_i^{GR} = \hat{U}_{\hat{R}_i}. \tag{7.4}$$

The GR estimates have an edf equal to $\hat{G}_k$ and ranks equal to the $\hat{R}_i$ and so are SEL-optimal for these features. Though there is no apparent attention to reducing coordinate-specific SEL, assigning the $\hat{U}$s to coordinates by a permutation vector z to minimize $\sum(\hat{U}_{z_i} - \eta_i)^2$ produces the same assignments as in (7.4). This equivalence indicates that the GR approach pays some attention to estimating the individual Bs and provides a structure for generalizing the approach to multivariate coordinate-specific parameters. As an added benefit of the GR approach, the $\hat{\theta}_i^{GR}$ are monotone transform equivariant, i.e., $[\widehat{h(\theta)}]^{GR} = h(\hat{\theta}_i^{GR})$ for monotone h(·).

7.1.2 *Triple goal estimates*

Since estimates of coordinate-specific parameters are inappropriate for producing histograms and ranks, no single set of values can simultaneously optimize all three goals. However, in many policy settings, communication and credibility is enhanced by reporting a single set of estimates with good performance for all three goals. To this end, Shen and Louis (1998) introduce "triple-goal" estimation criteria. Effective estimates are those that simultaneously produce a histogram that is a high quality estimate of the parameter histogram, induced ranks that are high quality estimates of the parameter ranks, and coordinate-specific parameter estimates that perform well. Shen and Louis (1998) evaluate and compare candidate estimates including the Ys, the PMs, CB, and GR estimates. Shen and Louis (1998) study properties of the various estimates, while Conlon and Louis (1999), Louis and Shen (1999), and Shen and Louis (2000) show how they work in practice. Here we summarize results.

As Shen and Louis (1998) show in the exchangeable case, the edf of the CB estimates is consistent for G if and only if the marginal distribution of η differs from G only by a location/scale change. The Gaussian/Gaussian model with θ the mean parameter qualifies, but in general CB estimates will not produce a consistent edf. However, $\bar{G}_k$ is consistent for G whenever G is identifiable from the marginal distributions (see Lindsay, 1983). Consistency is transform invariant.

PM dominates the other approaches for estimating individual θs under SEL, as it must. However, both GR and CB increase SEL by no more than 32% in all simulation cases, and both outperform use of the MLE (the Y_i Comparison of CB and GR is more complicated. As predicted by the Shen and Louis (1998) results, for the Gaussian-Gaussian case, CB and GR are asymptotically equivalent, while for the gamma-gamma case, model CB has a slightly better large-sample performance than does GR. However, for finite k, simulations show that GR can perform better than CB for skewed or long-tailed prior distributions.

As it must, GR dominates other approaches for estimating G_k and the advantage of GR increases with sample size. The Gaussian-Gaussian model provides the most favorable setup for CB relative to GR. Yet, the risk for CB exceeds GR by at least 39%. Performance for CB is worse in other situations, and in the Gaussian-lognormal case CB is even worse than PM. Importantly, the GR estimates are transform equivariant, and their performance in estimating G_k is transform invariant.

A start has been made on approaches to multiple-goal estimates for multivariate, coordinate-specific parameters. Ranks are not available in this situation, but coordinate-specific estimates based on the histogram are available. First, estimate G_k by a k-point discrete $\hat{G}_k$ that optimizes some histogram loss function. Then assign the mass points of $\hat{G}_k$ to coordinates so as to minimize SEL for the individual parameters. In the univariate parameter case, these estimates equal the GR estimates.

Finally, in addition to possible non-robustness to misspecification of the sampling likelihood (a potential problem shared by all methods), inferences from hierarchical models may not be robust to prior misspecification, especially if coordinate-specific estimates have relatively high variance. This non-robustness is most definitely a threat to the GR estimates, which depend on extreme percentiles of the posterior distribution. In the next subsection we discuss "robustifying" Bayes and empirical Bayes procedures.

7.1.3 Smoothing and robustness

Subsection 3.4.3 presented the EM algorithm approach to finding the nonparametric maximum likelihood (NPML) estimate of the prior distribution. Using it instead of a parametric prior (e.g., a conjugate) has been shown to be quite efficient and robust, especially for inferences based on only the first two moments of the posterior distribution (Butler and Louis, 1992). Since the NPML is always discrete with at most k mass points, smoothing is attractive for increased efficiency with small sample sizes. Approaches include directly smoothing the NPML, adding "pseudo-data" to the dataset and then computing the NPML, and model-based methods such as use of a Dirichlet process prior for G (see Section 2.5).

Laird and Louis (1991) propose an alternative, which they call *smoothing-by-roughening* (SBR), based on the EM algorithm method of finding the NPML. Shen and Louis (1999) provide additional evaluations. Though initiating the EM by a discrete prior with k mass points is best for producing the NPML, a continuous prior can initiate the EM. In this situation, the initiating prior $G^{(0)}$ and all subsequent iterates $(G^{(\nu)})$ are continuous. As $\nu \rightarrow \infty$, iterates converge very slowly to the NPML. Therefore. stopping the EM at iteration ν_k produces an estimate of the prior that is smoothed toward $G^{(0)}$ (equivalently, roughened away from $G^{(0)}$) with more smoothing for smaller values of ν_k. Note that the EM recursion uses the conditional expectation that produces $\bar{G}_k$. Specifically,

$$G^{(\nu+1)}(t) = E[G_k(t) \mid \mathbf{Y}, G^{(\nu)}] \,. \tag{7.5}$$

where $G^{(\nu)}$ plays the role of the prior.

As an example of smoothing-by-roughening and its relation to the NPML and ensemble estimates, Shen and Louis (1999) analyzed the baseball data from Efron and Morris (1973b). Batting averages are available for the first 45 at-bats of the 1970 season for 18 major league baseball players. The full-season batting average plays the role of the true parameter value. Batting averages are arc-sine transformed to stabilize their variance. and then these are centered around the mean of the transformed data. As a result, 0 is the center of the transformed observed batting average data.

The SBR starting distribution $G^{(0)}$ *is Uni f (-2.5.2.5).* and is approximated by 200 equi-probable, equally-spaced mass points. Figure 7.1 shows the cdf for the MMLE normal prior and the NPML. Notice that the NPML is concentrated at two points (i.e., it is very "rough"). The estimated normal prior is both smoothed and constrained to a narrow parametric form.

Figure 7.2 displays a sequence of roughened prior estimates $G^{(\nu)}$ along with the normal prior. Recall that the NPML is $G^{(\infty)}$. The sequence of estimates roughens from its *Uni f (-2.5,* 2.5) starting point toward the NPML. Convergence is extremely slow.

Finally, Figure 7.3 displays various choices for estimating G_k, including

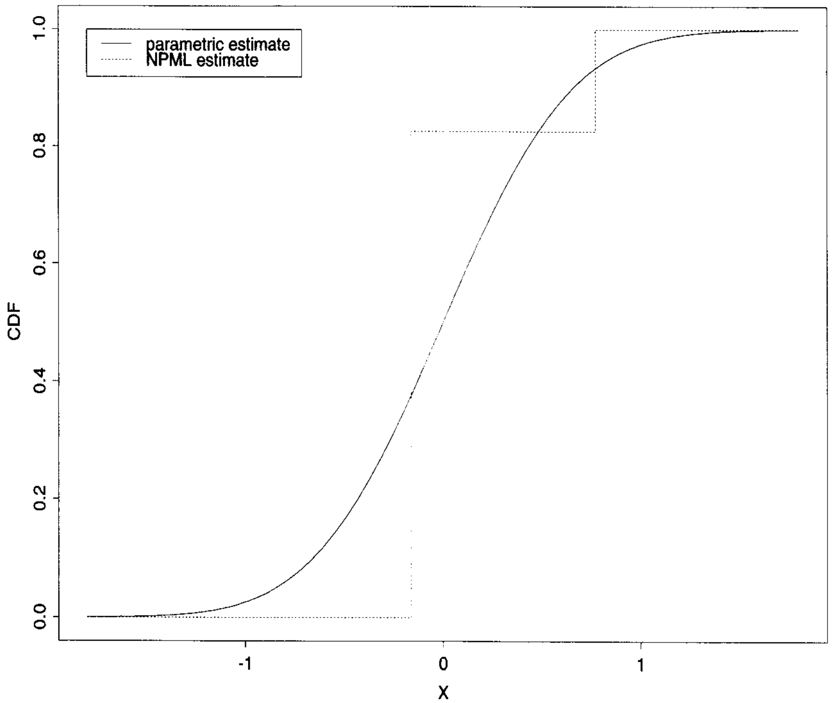


Figure 7.1 *Estimated parametric and nonparametric priors for the baseball data.*

the actual function (the edf of the transformed season batting averages). The remaining estimates (see the figure legend) are

- MLE: edf of the transformed, observed data
- PEB: the posterior means based on a MMLE normal prior
- NPEB: the posterior means based on the NPML prior
- CEB: the Louis and Ghosh constrained estimates
- SBR: $G^{(20)}$

Notice that the edf based on the MLE has a wider spread than that computed from the true parameter values. However, the posterior means from the PEB and the MPEB are shrunken too far toward the prior mean, and produce edfs with too small a spread. The CB approach gets the spread and shape about right, as does the SBR estimate from 20 iterations, though it produces a continuous distribution. Figure 3.1 presents another comparison of edfs based on the data (MLE), PEB and CB.

The SBR estimate is included in Figure 7.3 to show its alignment with

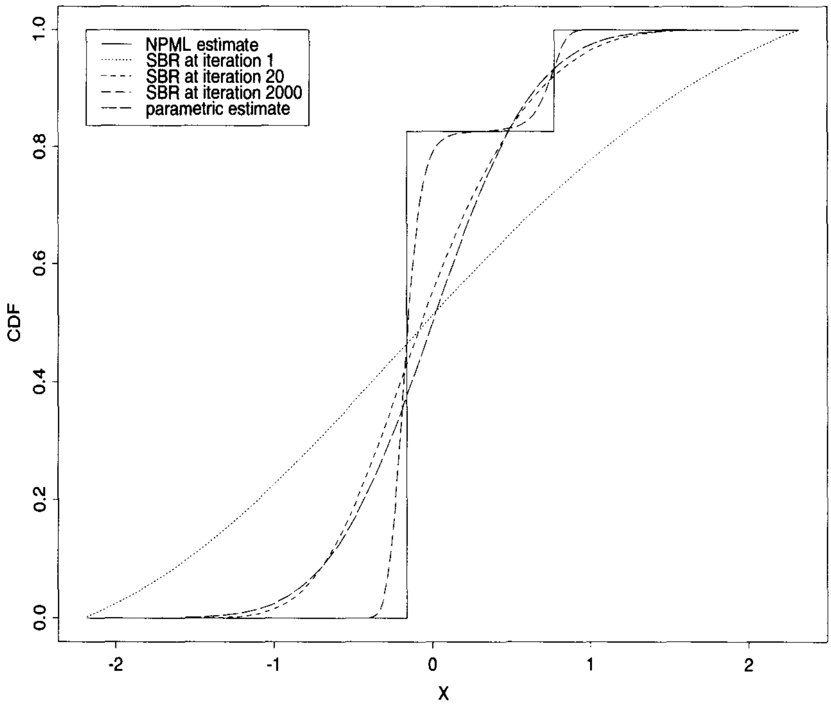


Figure 7.2 *Convergence of the smoothing-by-roughening (SBR) estimate from a uniform $G^{(0)}$ toward the NPML using the EM algorithm.*

the true G_k even though it is not a candidate (since it is continuous and the target distribution is discrete). A discretized version similar to $\hat{G}_k$ (see equation 7.2) may be effective, but further evaluation is needed.

The primary potential of the SBR estimate is as an estimate of G, not G_k. Such a smoothed estimate has the potential to produce robust and efficient (empirical) Bayes procedures. If $G^{(0)}$ is chosen to be the prior mean of a hyperprior, the SBR approach is similar to use of a Dirichlet process prior. Indeed, employing this idea for estimation of the mean in a normal/normal two-stage model, setting $\nu_k = a + b\log(k)$ exactly matches the estimate based on a parametric hyperprior. The parameters a and b capture the relative precision of the hyperprior and the data, and the number of iterates is always of order $\log(k)$.

The SBR and Dirichlet process approaches have the potential to produce procedures that are robust to prior misspecification and efficient for a family of priors. They are especially attractive when the number of sampling units

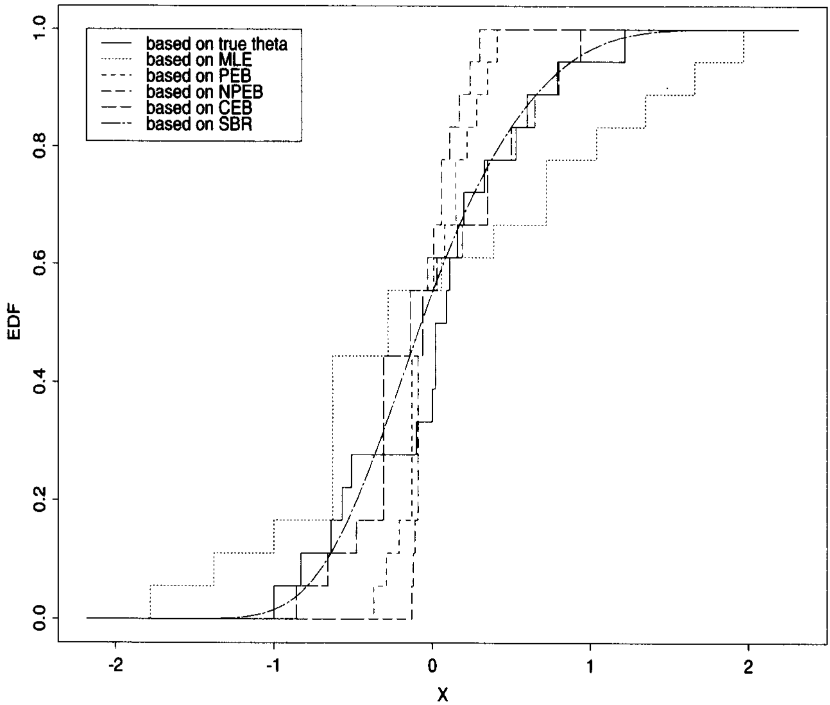


Figure 7.3 *Various estimates of G_k for the baseball data (see text for key to figure legend).*

k increases, but the information per unit does not (e.g., precision of the unit-specific MLEs stays constant). In these situations, if the parametric prior family is incorrect, procedures will have poor properties even for large k. Good performance for a broad range of loss functions will require that the SBR estimate (with ν_k iterations) or the posterior distribution from the Dirichlet process is consistent for G. The hyperprior Bayesian approach is, in principle, more attractive, since the posterior process incorporates residual uncertainty. Both approaches require further study.

7.2 Order restricted inference

Another area for which Bayes and empirical Bayes methods are particularly well suited is that of order restricted statistical inference. By this we mean modeling scenarios which feature constraints on either the sample space (truncated data problems) or the parameter space (constrained parameter problems). Such problems arise in many biostatistical settings, such as life table smoothing, bioassay, longitudinal analysis, and response surface modeling. While a great deal of frequentist literature exists on such problems (see for example the textbook by Robertson, Wright, and Dykstra, 1988), analysis is often difficult or even impossible since the order restrictions typically wreak havoc with the usual approaches for estimation and testing. By contrast, the Bayesian approach to analyzing such problems is virtually unchanged: we simply add the constraints to our Bayesian model (likelihood times prior) in the form of indicator functions, and compute the posterior distribution of the parameter of interest as usual. But again, this conceptual simplicity could not be realized in practice before the advent of sampling based methodology.

As an example, consider the problem of smoothing a series of raw mortality rates y_i into a gradually increasing final sequence θ_i, suitable for publication as a reference table. This process is referred to as "graduation" in the actuarial science literature. A restricted parameter space arises since the unknown $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$ must lie in the set S = $\{\boldsymbol{\theta} : \theta_1 < \theta_2 < \cdots < \theta_k\}$. Under the prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, we have that $p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})I_S(\boldsymbol{\theta})$ where $I_S(\boldsymbol{\theta})$ is the indicator function of the set S, so that $I_S(\boldsymbol{\theta})$ equals 1 when $\boldsymbol{\theta} \in S$, and equals 0 otherwise. Gelfand, Smith, and Lee (1992) point out that the complete conditional distribution for, say, θ_i satisfies

$$p(\theta_i|\mathbf{y}, \boldsymbol{\lambda}, \theta_{j\neq i}) \;\propto\; f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})I_{S_i}(\theta_i) \;, \tag{7.6}$$

where S_i is the appropriate cross-section of the constraint set S. In our increasing parameter case, for example, we have $S_i = \{\theta_i : \theta_{i-1} < \theta_i < \theta_{i+1}\}$. Hence if π is chosen to be conjugate with the likelihood f so that the unconstrained complete conditional distribution emerges as a familiar standard form, then the constrained version given in (7.6) is simply the same standard distribution restricted to S_i. Carlin (1992) and Carlin and Klugman (1993) use this technology to obtain increasing and increasing convex graduations of raw mortality data for males over age 30. These papers employ vague prior distributions in order to maintain objectivity in the results and keep the prior elicitation burden to a minimum.

Finally, Geyer and Thompson (1992) show how Monte Carlo methods may be used to advantage in computing maximum likelihood estimates in exponential families where closed forms for the likelihood function are unavailable. Gelfand and Carlin (1993) apply these ideas to arbitrary constrained and missing data models, where integration and sampling is over the data space, instead of the parameter space. These ideas could be applied to the problem of empirical Bayes point estimation in nonconjugate scenarios, where again the likelihood contains one or more intractable integrals.

7.3 Nonlinear models

As discussed in Chapter 3, the term "hierarchical model" refers to the extension of the usual Bayesian structure given in equation (2.1) to a hierarchy of 1 levels, where the joint distribution of the data and the parameters is given by

$$f(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}_3)\cdots\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\lambda}),$$

and we seek the marginal posterior of the first stage parameter $p(\boldsymbol{\theta}_1|\mathbf{y})$. In the case where f and the π_i are normal (Gaussian) distributions with known variance matrices, this marginal posterior is readily available using the results of Lindley and Smith (1972). However, more complicated settings generally require Monte Carlo methods for their solution. Gelfand et al. (1990) develop such techniques for arbitrary hierarchical normal linear models using the Gibbs sampler. Several biostatistical applications of such models have already appeared in the literature. Zeger and Karim (1991) apply the technology to a random effects model, and observe that high correlation between model parameters has a deleterious effect on algorithm convergence. Lange, Carlin, and Gelfand (1992) employ a hierarchical model for the longitudinal CD4 T-cell numbers of a cohort of San Francisco men infected with the HIV virus (see Sections 7.4 and 8.1 for an in-depth discussion of longitudinal data models). Their model allows for incomplete and unbalanced data, population covariates, heterogeneous variances, and errors-in-variables.

In many applications, a linear model is inappropriate for the underlying mean structure. Our reasons for rejecting the linear model may be theoretical (e.g., an underlying physical process) or practical (e.g., poor model fit). Fortunately, the Bayesian framework requires no major alterations to accommodate nonlinear mean structures. That is, our analytic framework is unchanged; only our computational algorithms must become more sophisticated (typically via Monte Carlo methods). The field of pharmacokinetics offers a large class of such models, since the underlying process of how a drug is processed by the body is known to follow a complex pattern that is nonlinear in the parameters (see e.g. Wakefield, 1996).

As another example, consider the problem of bioassay, i.e., the determination of an appropriate dose-response relationship from observed data. The example we shall use is modeling data from the Ames salmonella/microsome assay (Ames et al., 1973), perhaps the most widely used short term test for determining the mutagenic potency of certain chemical agents on strains of salmonella. The addition of a mutagenic agent to a plate of cultured microbes results in the growth of mutant colonies, clearly visible to the naked eye; the number of such colonies is an indicator of the mutagenic potential of the agent.

Krewski et al. (1993) review several biologically based models for Ames assay data. For a given agent-strain combination they suggest the following model for the expected number of mutant colonies μ as a function of dose:

$$\mu(D; \beta_0, \beta_1, \beta_2, \theta) = (\beta_0 + \beta_1 D) \exp\{-\beta_2 D^{\theta}\}, \tag{7.7}$$

where $\theta > 1$. The first factor reflects linear mutagenicity, while the second represents exponential toxicity. With mutagenic agents we typically have β_1 and β_2 positive, so that the tendency to mutate increases with dose, but so does the danger of cell death. The expected number of mutant colonies $\mu(D)$ generally increases with dose up to a certain point, after which it decreases, reflecting the effect of these two competing forces. Clearly $\beta_0 \geq 0$ since it is the expected number of mutant colonies at dose zero. The constraint $\theta > 1$ ensures that toxic effects are nonlinear and, in fact, negligible at low doses, so that the slope of the dose-response curve at $D = 0$, termed the *mutagenic potency,* is equal to β_1.

For a particular agent-strain combination, let X_{ikj} denote the observed number of mutant colonies observed at dose i ($i = 1, \ldots, I$) in the k-th laboratory ($k = 1, \ldots, K$) on replication j ($j = 1, \ldots, J$). A parametric Bayesian approach requires specification of the distribution of these observed counts. Krewski et al. (1993) point out that the observed counts may be expected to follow a Poisson distribution with some extra-Poisson variation. Rather than specify the form of this distribution, they assume that the variance of the observed counts V at any dose satisfies $V = \mu^p$, where μ is the expected number of mutant colonies at the same dose. Estimates of p and θ are obtained from the data by regression techniques. Quasi-likelihood estimation (McCullagh and Nelder, 1989, Chapter 9) is subsequently employed to estimate the β_i while keeping p and θ fixed at their estimated values.

We take account of extra-Poisson variability by allowing the Poisson parameter to vary randomly among replications and/or laboratories. A possible three-stage Bayesian hierarchical model is

$$\begin{array}{rcl} X_{ikj} & \stackrel{ind}{\sim} & Po(\lambda_{ik}), \\ \text{where} \quad \lambda_{ik} & = & \mu(D_i; \beta_0, \beta_{1k}, \beta_2, \theta); \\ \log(\beta_{1k}) & \stackrel{iid}{\sim} & N(\eta, \tau^2), \end{array} \tag{7.8}$$

where Po denotes the Poisson distribution and N denotes the normal (Gaussian) distribution. Normality of the log mutagenic potencies across laboratories was suggested by Myers et al. (1987). Prior distributions on θ, β_0, and β_2 add to the model's second stage, while the priors on η and τ^2 form the third stage, completing the hierarchy. Although mutagenic potency is the quantity of primary interest, the models for β_0 and β_2 should perhaps be similar to that for β_1. As such, model (7.8) might be modified to incorporate random β_{0l} and β_{2k} effects with little increase in conceptual and computational complexity.

Model (7.8) allows explicit estimation of inter-laboratory variability in mutagenic potency. But it does not permit estimation of intea-laboratory variability, and thus its comparison with inter-laboratory variability. In order to estimate these variance components, we can use a four-stage hierarchical model of the form

$$\begin{array}{rcl} X_{ikj} & \stackrel{ind}{\sim} & Po(\lambda_{ikj}), \\ \text{where} \;\; \lambda_{ikj} & = & \mu(D_i; \beta_0, \beta_{1kj}, \beta_2, \theta); \\ \log(\beta_{1kj}) & \stackrel{ind}{\sim} & N(\eta_k, \tau^2); \\ \eta_k & \stackrel{iid}{\sim} & N(\mu, \sigma^2) \,. \end{array} \tag{7.9}$$

In this case the intra-laboratory variability is quantified by τ^2, while the inter-laboratory variability is quantified by σ^2. Concerns over the possible non-normality of the distributions of the η_k are mitigated by the fact that distributional changes at the third and higher stages of a hierarchy typically have limited impact on posterior distributions of second stage parameters (see Berger, 1985, p. 108). Identifying prior distributions on θ, β_0, β_2, μ, τ^2, and σ^2 completes the Bayesian model specification.

A principal purpose of many bioassays is to determine an agent's suitability as a reference standard. However, all of the models presented thus far are capable of handling only a single chemical, rather than both a new agent and a reference agent. When experimental observations on a known standard agent are available, we can extend model (7.8) to

$$\begin{array}{rcl} X_{ikj}^{(u)} & \stackrel{ind}{\sim} & Po(\lambda_{ik}^{(u)}), \text{ where} \\ \lambda_{ik}^{(u)} & = & \mu(D_i; \beta_0^{(u)}, \beta_{1k}^{(u)}, \beta_2^{(u)}, \theta^{(u)}); \\ \log(\beta_{1k}^{(u)}) & \stackrel{iid}{\sim} & N(\eta^{(u)}, (\tau^{(u)})^2), \end{array}$$

and

$$\begin{array}{rcl} X_{ikj}^{(s)} & \stackrel{ind}{\sim} & Po(\lambda_{ik}^{(s)}), \text{ where} \\ \lambda_{ik}^{(s)} & = & \mu(D_i; \beta_0^{(s)}, \beta_{1k}^{(s)}, \beta_2^{(s)}, \theta^{(s)}); \\ \log(\beta_{1k}^{(s)}) & \stackrel{iid}{\sim} & N(\eta^{(s)}, (\tau^{(s)})^2), \end{array}$$

where a (u) superscript denotes the unknown agent while an (s) denotes the standard agent. Then the posterior distribution of $\gamma_k \equiv \beta_{1k}^{(u)} / \beta_{1k}^{(s)}$ summarizes the relative mutagenic potency of the unknown agent to that of the standard for lab k, $k = 1, \ldots, K$. If all K of these posterior distributions are similar, one can conclude that the unknown agent is a useful reference standard.

In a preliminary analysis of a single International Program on Chemical Safety (IPCS) agent-strain combination (urban air particles and strain TA100-S9), Etzioni and Carlin (1993) assume independence among the four prior components and adopt familiar distributional forms, such as a normal density for M and gamma densities for the positive parameters V, β_0, and β_2. They then obtain posterior means and standard deviations for β_0, β_2, M, and V under the three-stage model (7.8). While this model requires only elementary quadrature methods, it is too basic to answer all of the authors' research questions - most notably the intra- versus inter-lab variability question. Still, their Bayesian analysis contrasts markedly with the method of Krewski et al. (1993), who addressed these two estimation goals for the IPCS data in two separate steps. In the latter approach, for a given agent-strain combination a separate estimate of mutagenic potency was obtained for each repetition in each laboratory. Then analysis of variance was used to decompose the variance of the potency estimates into within and between laboratory components.

The more complicated nonlinear hierarchical models described above require careful application of Monte Carlo methods. For example, in model (7.9) it is not possible to write the loglikelihood as a sum of low dimensional integrals. More explicitly, here we would define $\boldsymbol{\beta} = \{\beta_{1kj}\}$, a vector of length KJ. If we could obtain closed forms for the integrals over the $\{\eta_k\}$, we would still have that β_{1kj} and $\beta_{1kj'}$ are marginally dependent given μ and σ^2. Hence the likelihood would now at best be a product of J-dimensional integrals. Of course, in the case of hierarchical normal linear models, such integration is still feasible largely in closed form, thanks to the well-known work of Lindley and Smith (1972). In our heavily nonconjugate scenario, however, no such convenient forms are available and, due to the model's high dimensionality, we would likely turn to Monte Carlo integration methods. Here, the nonconjugate structure present at the first two stages of model (7.9) precludes closed forms for the complete conditional distributions of the β_{1kj}. Generation of these values would have to proceed via some sort of Metropolis-Hastings or other rejection algorithm (e.g., Gilks and Wild, 1992).

7.4 Longitudinal data models

A great many datasets arising in statistical and biostatistical practice consist of repeated measurements (usually ordered in time) on a collection of individuals. Data of this type are referred to as *longitudinal data,* and require special methods for handling the correlations that are typically present among the observations on a given individual. More precisely, suppose we let Y_{ij} denote the measurement on the i^{th} individual in the study, $j = 1, \ldots, s_i$ and $i = 1, \ldots, n$. Arranging each individual's collection of observations in a vector $Y_i = (Y_{i1}, \ldots, Y_{is_i})^T$, we can fit the usual fixed effects model

$$Y_i = \mathbf{X}_i \alpha + \epsilon_i \ ,$$

where $\mathbf{X}_i$ is an $s_i \times p$ design matrix of covariates and a is a $p \times 1$ parameter vector. While the usual assumptions of normality, zero mean, and common variance might still be appropriate for the components of ϵ_i, generally they are correlated with covariance matrix $\boldsymbol{\Sigma}$. This matrix can be general or follow some simplifying pattern, such as compound symmetry or a first-order autoregression. Another difficulty is that data frequently exhibit more variability than can be adequately explained using a simple fixed effects model.

A popular alternative is provided by the *mixed* model

$$Y_i = \mathbf{X}_i\alpha + \mathbf{W}_i\beta_i + \epsilon_i \ , \tag{7.10}$$

where $\mathbf{W}_i$ is an $s_i \times q$ design matrix (q typically less than p), and β_i is a $q \times 1$ vector of subject-specific random effects, usually assumed to be normally distributed with mean vector 0 and covariance matrix V. The β_is capture any subject-specific mean effects, and also enable the model to accurately reflect any extra variability in the data. Furthermore, it may now be realistic to assume that, given β_i, the components of ϵ_i *are* independent. This allows us to set $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_{s_i}$ (as in the original paper by Laird and Ware, 1982), although an autoregressive or other structure for $\boldsymbol{\Sigma}$ may still be appropriate. Note that the marginal covariance matrix for Y_i includes terms induced both by the random effects (from V) and the likelihood (from $\boldsymbol{\Sigma}$).

The model (7.10) can accommodate the case where not every individual has the same number of observations (unequal s_i), perhaps due to missed office visits, data transmission errors, or study withdrawal. In fact, features far broader than those suggested above are possible in principle. For example, heterogeneous error variances σ_i^2 may be employed, removing the assumption of a common variance shared by all subjects. The Gaussian errors assumption may be dropped in favor of other symmetric but heavier-tailed densities, such as the Student's t. Covariates whose values change and growth curves that are nonlinear over time are also possible within this framework.

An increasing number of commercially available computer packages (e.g., SAS **`Proc Mixed`** and BMDP **`5V`**) enable likelihood analysis for mixed models; see Appendix Section C.2 for a discussion. Though these packages employ sophisticated numerical methods to estimate the fixed effects and covariance matrices, in making inferences they are unable to include the variance inflation and other effects of having estimated the features of the prior distributions. By contrast, MCMC methods are ideally suited to this goal, since they are able to perform the necessary integrations in our high-dimensional (and typically analytically intractable) hierarchical model. Several papers (e.g., Zeger and Karim, 1991; Lange et al., 1992; Gilks, Wang et al., 1993; McNeil and Gore, 1996) have employed MCMC methods in longitudinal modeling settings. Chib and Carlin (1999) provide a method for generating the fixed effects α and all the random effects β_i in a single block, which greatly improves the convergence of the algorithm.

We require prior distributions for α, σ^2, and V to complete our model specification. Bayesians often think of the distribution on the random effects β_i as being part of the prior (the "first stage," or *structural* portion), with the distribution on $\mathbf{V}$ forming another part (the "second stage," or *subjective* portion). Together, these two pieces are sometimes referred to as a *hierarchical prior.* Again, the case study in Section 8.1 provides a real-data illustration.

7.5 Continuous and categorical time series

Many datasets arising in statistics and biostatistics exhibit autocorrelation over time that cannot be ignored in their analysis. A useful model in such situations is the *state space model,* given in its most general form by

$$\begin{array}{rl} x_t = f_t(x_{t-1}) + u_t, & \text{and} \\ y_t = h_t(x_t) + v_t, & t = 1, \ldots, n, \end{array} \tag{7.11}$$

where x_t is the (unobserved) $p \times 1$ state vector, y_t is the (observed) $q \times 1$ observation vector, and $f_t(\cdot)$ and $h_t(\cdot)$ are known, possibly nonlinear functions which depend on some unknown parameters. Typically, u_t and v_t are taken as independent and identically distributed, with $u_t \sim N_p(0, \Sigma)$ and $v_t \sim N_q(0, \Upsilon)$, where N_p denotes the p-dimensional normal distribution.

State space models are particularly amenable to a Bayesian approach, since the time-ordered arrival of the data means that the notion of updating prior knowledge in the presence of new data arises quite naturally. Carlin, Polson, and Stoffer (1992) use Monte Carlo methods in an analysis of these models that does not resort to convenient normality or linearity assumptions at the expense of model adequacy. Writing the collection of unknown parameters simply as $\boldsymbol{\theta}$ for the moment, the likelihood for model (7.11) is given by

$$L(\boldsymbol{\theta}|x_0, \mathbf{x}, \mathbf{y}) = p(x_0) \prod_{t=1}^{n} g_u(x_t|x_{t-1}, \boldsymbol{\theta}) \prod_{t=1}^{n} g_v(y_t|x_t, \boldsymbol{\theta}) \tag{7.12}$$

for some densities g_u and g_v. Next, introducing the nuisance parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ they assume that

$$\begin{array}{rl} p(x_t|x_{t-1}, \lambda_t, \Sigma) = N_p(f_t(x_{t-1}), \lambda_t \Sigma), & \text{and} \\ p(y_t|x_t, \omega_t, \Upsilon) = N_q(h_t(x_t), \omega_t \Upsilon), & t = 1, \ldots, n \ , \end{array} \tag{7.13}$$

where Σ and Υ are parameter matrices of sizes $p \times p$ and $q \times q$, respectively. Hence unconditionally,

$$\begin{array}{rl} g_u(x_t|x_{t-1}, \Sigma) = \int_\Lambda p(x_t|x_{t-1}, \lambda_t, \Sigma) p_1(\lambda_t) d\lambda_t, & \text{and} \\ g_v(y_t|x_t, \Upsilon) = \int_\Omega p(y_t|x_t, \omega_t, \Upsilon) p_2(\omega_t) d\omega_t, & t = 1, \ldots, n \ . \end{array} \tag{7.14}$$

Note that, by varying $p_1(\lambda_t)$ and $p_2(\omega_t)$, the distributions g_u and g_v are scale mixtures of multivariate normals for each t, thus enabling a wide variety of nonnormal error densities to emerge in (7.14). For example, in the univariate case, the distributions g_u and g_v can be double exponential, logistic, exponential power, or *t* densities.

As discussed just prior to Example 6.2, the conditioning feature of the Gibbs sampler makes such normal scale mixtures very convenient computationally. Non-Gaussian likelihoods may be paired with non-Gaussian priors and Gaussian complete conditionals may still emerge for the state parameters (in the present case, the x_ts). Further, under conjugate priors the complete conditionals for Σ and Υ follow from standard Gaussian and Wishart distribution theory, again due to the conditioning on $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$. Finally, the complete conditionals for the nuisance parameters themselves are of known functional form. For example, for our state space model we have $p(\lambda_t|\lambda_{j\neq t}, \boldsymbol{\omega}, \Sigma, \Upsilon, \mathbf{y}, \mathbf{x}) = p(\lambda_t|\Sigma, x_t, x_{t-1}) \propto p(x_t|x_{t-1}, \lambda_t, \Sigma)p_1(\lambda_t)$ by Bayes' Theorem. But the normalizing constant is known by construction: it is $g_u(x_t|x_{t-1}, \Sigma)$ from equation (7.14). Complete conditionals for the ω_t would of course arise similarly.

Thus, the sampling based methodology allows the solution of several previously intractable modeling problems, including the modeling of asymmetric densities on the positive real line (as might be appropriate for death rates), explicit incorporation of covariates (such as age and sex), heteroscedasticity of errors over time, multivariate analysis (including simultaneous modeling of both preliminary and final mortality estimates), and formal model choice criteria for selecting the best model from many. West (1992) uses importance sampling and adaptive kernel-type density estimation for estimation and forecasting in dynamic nonlinear state space models. In the special case of linear state space models, Carter and Kohn (1994) provide an algorithm for generating all of the elements in the state vector simultaneously, obtaining a far less autocorrelated sampling chain (and hence faster convergence) than would be possible using the fully conditional algorithm of Carlin, Polson, and Stoffer (1992). See also Carter and Kohn (1996) for further improvements in the context of conditionally Gaussian state space models. Chib and Greenberg (1995b) extend this framework to allow time invariant parameters in the state evolution equation, as well as multivariate observations in a vector autoregressive or moving average process of order l. Marriott et al. (1996) use a hybrid Gibbs-Metropolis algorithm to analyze the exact ARMA likelihood (rather than that conditional on initial conditions), and incorporate the associated constraints on the model parameters to insure stationarity and invertibility of the process.

Carlin and Polson (1992) extend model (7.11) further, formulating it for categorical data by adding the condition $y_t = A(y_t^*)$ to the specification, where A(·) is a given many-to-one function. They also provide an example of biostatistical interest, applying the methodology to the infant sleep state study described by Stoffer et al. (1988). In this dataset, the experimenter observes only $y_t = 1$ or 0, corresponding to REM or non-REM sleep, respectively, and the goal is to estimate the infant's underlying (continuous) sleep state x_t. Simple density estimates for the x_t are available both prior to the observation of y_t (prediction) and afterward (filtering). Albert and Chib (1993b) use a similar approach to handle underlying autoregressive structure of any order, and use the resulting hidden Markov model to advantage in analyzing U.S. GNP data. Chib (1996) provides a multivariate updating algorithm for a collection of discrete state variables, analogous to the result of Carter and Kohn (1994) for continuous variables in a state space model. See also Chib and Greenberg (1998) for MCMC Bayesian methods for analyzing multivariate probit models, even more general approaches for correlated binary response data.

7.6 Survival analysis and frailty models

Bayesian methods have been a subject of increasing interest among researchers engaged in the analysis of survival data. For example, in monitoring the results of a clinical trial, the Bayesian approach is attractive from the standpoint of allowing the input of additional relevant information concerning stopping or continuing the trial. Spiegelhalter et al. (2000) provide a remarkably comprehensive overview of and guide to the literature on Bayesian methods in clinical trials and other areas of health technology assessment (e.g., observational studies). Areas covered include prior determination, design, monitoring, subset analysis. and multicenter analysis; a guide to relevant websites and software is also provided.

However, the Bayesian approach may be somewhat limited in its ability to persuade a broad audience (whose beliefs may or may not be characterized by the selected prior distribution) as to the validity of its results. In an attempt to remedy this, Chaloner et al. (1993) developed software for eliciting prior beliefs and subsequently demonstrate its application to a collection of P relevant parties (medical professionals, policy makers, patients, etc.) in an attempt to robustify the procedure against outlying prior opinions. Modern computational methods allow real-time evaluation of the trial's stopping rule, despite the P-fold increase in the computational burden.

7.6.1 Statistical models

We now outline the essential features of a survival model. Suppose that associated with individual i in our study, $i = 1, \ldots, n$, is a p-dimensional covariate vector $\mathbf{z}_i$ which does not change over time. We model the effect of these covariates using the familiar log-linear proportional hazards form, namely $h(t|\mathbf{z}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{z})$, where $\boldsymbol{\beta}$ is a p-vector of unknown parameters. Suppose that d failures are observed in the combined group over the course of the study. After listing the individuals in order of increasing time to failure or censoring, let $\mathbf{z}_i$ be the covariate vector associated with the i^{th}

individual in this list, $i = 1, \ldots, n$. Denote the j^{th} *risk set,* or the collection of individuals still alive at the time of the j^{th} death, by $\mathcal{R}_j$ for $j = 1, \ldots, d$. Then the Cox partial likelihood (Cox and Oakes, 1984, Chapter 7) for this model is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} \left(\frac{e^{\boldsymbol{\beta}^T \mathbf{z}_j}}{\sum_{k \in \mathcal{R}_j} e^{\boldsymbol{\beta}^T \mathbf{z}_k}} \right) , \tag{7.15}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\mathbf{z}_j = (z_{1j}, \ldots, z_{pj})$. It is often convenient to treat this partial likelihood as a likelihood for the purpose of computing the posterior density. A justification for this is offered by Kalbfleisch (1978), who places a gamma process prior on the baseline hazard function (independent of the prior for $\boldsymbol{\beta}$), and shows that the marginal posterior density for $\boldsymbol{\beta}$ approaches a form proportional to the Cox partial likelihood as the baseline hazard prior becomes arbitrarily diffuse.

To complete the Bayesian model we must specify a joint prior distribution $\pi(\boldsymbol{\beta})$ on the components of $\boldsymbol{\beta}$. Before doing so, however, we focus attention more clearly on the monitoring of treatment-control clinical trials. Suppose the variable of primary interest is the first component of **z**, z_1 a binary variable indicating whether individual *i* has been randomly assigned to the control group ($z_1 = 0$) or the treatment group ($z_1 = 1$). Hence interest focuses on the parameter β_1 and in particular on whether $\beta_1 = 0$. Now returning to the question of prior elicitation for $\boldsymbol{\beta}$, we make the simplifying assumption that $\pi(\boldsymbol{\beta}) = \pi_1(\beta_1)\pi_{-1}(\boldsymbol{\beta}_{-1})$, where $\boldsymbol{\beta}_{-1} = (\beta_2, \ldots, \beta_p)$. That is, β_1 is *a priori* independent of the covariate-effect parameters $\boldsymbol{\beta}_{-1}$. In the absence of compelling evidence to the contrary, this *a priori* lack of interaction between the treatment effect and covariates provides a reasonable initial model. Moreover, this prior specification does not force independence in the posterior, where the data determine the appropriate level of dependence. In addition, we assume a flat prior on the nuisance parameters, namely $\pi_{-1}(\boldsymbol{\beta}_{-1}) \equiv 1$. This eliminates specifying the exact shape of the prior, and produces an evaluation of covariate effects comparable to that produced by traditional frequentist methods. Of course, if objective evidence on covariate effects is available, it can be used. However, since "standard practice" in patient care changes rapidly over time, such a prior should be fairly vague or include a between-study variance component.

7.6.2 Treatment effect prior determination

It remains to specify the (univariate) prior distribution on the treatment effect, $\pi_1(\beta_1)$. An algorithm and corresponding software for eliciting such a prior from an expert was developed by Chaloner et al. (1993); see Appendix Section C.1 for details. While this method has been an operational boon to the Bayesian approach to this problem, its use of subjective expert opinion as the sole source of evidence in creating the prior distribution on β_1 may constitute an unnecessary risk in many trials. Objective development of a prior for the control-group failure rate may be possible using existing data, since this value depends only on the endpoint selected, and not the particular treatment under investigation. Even so, eligibility, diagnostic and other temporal drifts limit the precision of such a prior. Data relevant for the treatment group are less likely to exist, but may be available in the form of previous randomized trials of the treatment. Failing this, existing non-randomized trial data might be employed, though prior uncertainty should be further inflated here to account for the inevitable biases in such data.

A sensible overall strategy for prior determination in clinical trials is recommended by Spiegelhalter, Freedman, and Parmar (1994). These authors suggest investigating and reporting the posterior distributions that arise over a *range* of priors, in the spirit of sensitivity analysis (see Subsection 6.1.1). For example, we might investigate the impact of a "clinical" prior, which represents the (typically optimistic) prior feelings of the trial's investigators, a "skeptical" prior, which reflects the opinion of a person or regulatory agency that doubts the treatment's effectiveness, and a "non-informative" prior, a neutral position that leads to posterior summaries formally equivalent to those produced by standard maximum likelihood techniques. The specific role of the prior and the uncertainty remaining in the results after the data have been observed can then be directly ascertained. Appendix Subsection C.3.1 contains a brief description of **iBART**, a software program for implementing this approach with a normal or other survival likelihood for the treatment effect parameter.

7.6.3 Computation and advanced models

For the above models and corresponding priors, we require efficient algorithms for computing posterior summaries (means, variances, density estimates, etc.) of the quantities of interest to the members of the clinical trial's data monitoring board. Naturally β_1 itself would be an example of such a quantity; the board might also be interested in the Bayes factor BF in favor of $H_0 : \beta_1 \neq 0$, or perhaps simply the posterior probability that the treatment is efficacious, given that this has been suitably defined. For simplicity of presentation, in what follows we concentrate solely on the posterior of β_1. The simplest method for summarizing β_1's posterior would be to use the first order (normal) approximation provided by Theorem 5.1. Such an approximation is extremely attractive in that it is computationally easy to produce quantiles of the posterior distribution, as well as a standardized Z-score for testing $H_0 : \beta_1 = \beta_1^{(0)}$ (for example, $Z = (E(\beta_1|\mathcal{R}) - \beta_1^{(0)})/\sqrt{V(\beta_1|\mathcal{R})}$). This statistic allows sequential graphical monitoring displays similar to traditional frequentist monitoring plots (e.g. Lan and DeMets, 1983).

If n is not sufficiently large, results appealing to asymptotic posterior normality will be inappropriate, and we must turn to more accurate methods such as Laplace's method or traditional numerical quadrature, such as the Gauss-Hermite rules. In fact, if the dimension of $\boldsymbol{\beta}$ is small, it will be very convenient to simply discretize the entire vector (instead of just β_1) and use a naive summation method, such as the trapezoidal rule. For high-dimensional $\boldsymbol{\beta}$, Monte Carlo integration methods will likely provide the only feasible alternative.

These methods are especially convenient for an extension of model (7.15) known as a frailty *model.* Introduced by Vaupel, Manton, and Stallard (1979), this is a mixed model with random effects (called frailties) that correspond to an individual's health status. That is, we replace the Cox hazard function for subject i, $h(t|\mathbf{z}_i, \boldsymbol{\beta}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{z}_i)$, by

$$h(t|\mathbf{z}_i, \boldsymbol{\beta}, \gamma_i) = h_0(t)\,\gamma_i\,\exp(\boldsymbol{\beta}^T\mathbf{z}_i)\ ,$$

where the frailties γ_i are typically assumed to be i.i.d. variates from a specified distribution (e.g., gamma or lognormal). Clayton (1991) and Sinha (1993) show how the Gibbs sampler may be used to fit this model under an independent increment gamma process parametrization of the baseline hazard function. Gustafson (1994) and Sargent (1995) avoid this assumption by treating the Cox partial likelihood as a likelihood, and employing more sophisticated MCMC approaches.

A nonnormal posterior distribution for β_1 precludes traditional monitoring plots based on standardized Z-scores. However, Bayesian monitoring may still proceed by plotting the posterior probabilities that β_1 falls outside an "indifference zone" (or "range of equivalence"), $(\beta_{1,L}, \beta_{1,U})$. Recall from Subsection 6.1.2 that this approach, proposed by Freedman and Spiegelhalter (1983, 1989, 1992), assumes that for β_1 values within this prespecified range, we are indifferent as to the practical use of either the treatment or the control preparation. Since negative values of β_1 indicate an effective treatment, we can set $\beta_{1,L} = K < 0$ and $\beta_{1,U} = 0$, an additional benefit perhaps being required of the treatment in order to justify its higher cost in terms of resources, clinical effort, or toxicity. The Bayesian decision rule terminates the trial (or more realistically, provides a guideline for termination) when $P(\beta_1 > \beta_{1,U}|\mathcal{R})$ is sufficiently small (deciding in favor of the treatment), or when $P(\beta_1 < \beta_{1,L}|\mathcal{R})$ is sufficiently small (deciding in favor of the control). The case study in Section 8.2 provides an illustration.

7.7 Sequential analysis

As illustrated in equation (2.4), when data samples $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K$ arrive over time, Bayes' Rule may be used sequentially simply by treating the current posterior as the prior for the next update, i.e.,

$$p(\boldsymbol{\theta}|\mathbf{y}_1,\ldots,\mathbf{y}_K) \propto f_K(\mathbf{y}_K|\boldsymbol{\theta},\mathbf{y}_1,\ldots,\mathbf{y}_{K-1})p(\boldsymbol{\theta}|\mathbf{y}_1,\ldots,\mathbf{y}_{K-1}) \, .$$

This formula is conceptually straightforward and simple to implement provided all that is required is an updated posterior for θ at each monitoring point. In many cases, however, the analyst faces a specific *decision* at each timepoint - say, a repeated hypothesis test regarding θ, the outcome of which may determine whether the experiment should be continued or not. For the frequentist, this raises the thorny issues of repeated significance testing and multiple comparisons. To protect the overall error rate α, many classical approaches resort to "α-spending functions". as in the aforementioned work by Lan and DeMets (1983).

In the Bayesian framework, the problem is typically handled by combining the formal decision theoretic tools of Appendix B with the usual Bayesian probability tools. The traditional method of attack for such problems is *backward induction* (see e.g. DeGroot, 1970, Chapter 12). a method which, while conceptually simple, is often difficult to implement when K is at all large. Subsection 7.7.1 describes a class of two-sided sequential decision problems (i.e., their likelihood, prior, and loss structures). while Subsection 7.7.2 outlines their solution via the traditional backward induction approach. Then in Subsection 7.7.3, we present a *forward* sampling algorithm which can produce the optimal stopping boundaries for a broad subclass of these problems at far less computational and analytical expense. In addition, this forward sampling method can *always* be used to find the best boundaries within a certain class of decision rules, even when backward induction is infeasible. We then offer a brief discussion, but defer exemplification of the methods until Subsection 8.2.4, as part of our second Chapter 8 case study.

7.7.1 Model and loss structure

Carlin, Kadane, and Gelfand (1998) present the sequential decision problem in the context of clinical trial monitoring. Since this setting is very common in practice, and is one we have already begun to develop in Sections 6.1.2 and 7.6, we adopt it here as well.

Suppose then that a clinical trial has been established to estimate the effect θ of a certain treatment. Information already available concerning 0 is summarized by a prior distribution $p(\theta)$. As patient accrual and followup is getting underway, the trial's data safety and monitoring board (hereafter referred to simply as "the board") agrees to meet on K prespecified dates to discuss the trial's progress and determine whether or not it should be continued. At each monitoring point $k \in \{1, \ldots, K\}$, some cost A_k will have been invested in obtaining a new datapoint y_k having distribution $f(y_k|\theta)$, enabling computation of the updated posterior distribution $p(\theta|y_1,\ldots,y_k)$.

It is convenient to think of $k = 0$ as an additional, preliminary monitoring point, allowing the board to determine whether the trial should be run at all, given only the prior and the set of costs $\{A_k\}$.

Suppose that for the treatment effect parameter $\theta \in \Re$, large negative values suggest superiority of the treatment, while large positive values suggest superiority of the placebo (inferiority of the treatment). Suppose further that in conjunction with substantive information and clinical expert opinion, an indifference zone (θ_L, θ_U) can be determined such that the treatment is preferred for $\theta < \theta_L$ the placebo for $\theta > \theta_U$ and either decision is acceptable for $\theta_L < \theta < \theta_U$.

At the final monitoring point (K), there are then two decisions available to the board: d_1 (stop and decide in favor of the treatment), or d_2 (stop and decide in favor of the placebo). Corresponding to these two decisions, the indifference zone (θ_L, θ_U) suggests the following loss functions for given true values of $\theta \in \Re$:

$$l_1^{(K)}(d_1, \theta) = s_1^{(K)}(\theta - \theta_U)^+ \ , \ \text{and} \ \ l_2^{(K)}(d_2, \theta) = s_2^{(K)}(\theta_L - \theta)^+ \ , \qquad (7.16)$$

where $s_1^{(K)}, s_2^{(K)} > 0$ and the "+" superscript denotes the "positive part" function, i.e., $y^+ \equiv \max(0, y)$. That is, incorrect decisions are penalized, with the penalty increasing linearly as θ moves further from the indifference zone. To facilitate the presentation of backward induction later in this section, for the moment we simplify the loss functions (7.16) by dropping the positive part aspect, resulting in linearly increasing *negative* losses (gains) for correct decisions.

Given the posterior distribution of the treatment parameter at this final time point, $p(\theta|y_K)$, the best decision is that corresponding to the smaller of the two posterior expected losses, $L_1^{(K)} \equiv E[l_1^{(K)}(d_1, \theta)|y_K] = s_1^{(K)}[E(\theta|y_K) - \theta_U]$, and $L_2^{(K)} \equiv E[l_2^{(K)}(d_2, \theta)|y_K] = s_2^{(K)}[\theta_L - E(\theta|y_K)]$. If $E(\theta|y_K)$ exists, then clearly $L_1^{(K)}$ increases in $E(\theta|y_K)$ while $L_2^{(K)}$ decreases in $E(\theta|y_K)$. The breakpoint γ_K between the optimality of the two decisions follows from equating these two expected losses. Specifically, decision d_1 (decide in favor of the treatment) will be optimal if and only if

$$E(\theta|y_K) \leq \gamma_K \equiv \frac{s_1^{(K)}\theta_U + s_2^{(K)}\theta_L}{s_1^{(K)} + s_2^{(K)}} \qquad (7.17)$$

otherwise (when $E(\theta|y_K) > \gamma_K$), decision d_2 is optimal. Thus when $s_1^{(K)} = s_2^{(K)}$ the breakpoint γ_K is simply the midpoint of the indifference zone. A similar result holds under the positive part loss functions in (7.16).

7.7.2 Backward induction

To obtain the optimal Bayesian decision in the sequential setting at any time point other than the last, one must consider the probability of data values that have not yet been observed. This requires "working the problem backward" (from time K back to time 0), or backward induction. To illustrate, consider time K - 1. There are now three available decisions: d_1, d_2, and d_3 (continue and decide after y_K is available for analysis). To simplify the notation, suppose for now that $K = 1$, so that no data have yet accumulated and hence the current posterior $p(\theta|y_{K-1})$ equals the prior $p(\theta)$. The posterior expected losses are now $L_1^{(K-1)} = s_1^{(K-1)}[E(\theta) - \theta_U]$, $L_2^{(K-1)} = s_2^{(K-1)}[\theta_L - E(\theta)]$, and

$$
\begin{aligned}
L_3^{(K-1)} = A_K &+ \int_{R_1^{(K)}} \int_{-\infty}^{\infty} s_1^{(K)}(\theta - \theta_U)p(\theta|y_K)m(y_K)d\theta dy_K \\
&+ \int_{R_2^{(K)}} \int_{-\infty}^{\infty} s_2^{(K)}(\theta_L - \theta)p(\theta|y_K)m(y_K)d\theta dy_K \ , \qquad (7.18)
\end{aligned}
$$

where $R_1^{(K)} = \{y_K : E(\theta|y_K) \le \gamma_K\}$, $R_2^{(K)} = \{y_K : E(\theta|y_K) > \gamma_K\}$, γ_K is as given in (7.17), and $m(y_K)$ is the marginal density of the (as yet unobserved) data value y_K Once the prior $p(\theta)$ and the sampling cost A_K are chosen, the optimal decision is that corresponding to the minimum of $L_1^{(K-1)}$, $L_2^{(K-1)}$ and $L_3^{(K-1)}$. Alternatively, a class of priors (say, the normal distributions having fixed variance σ^2) might be chosen, and the resulting $L_i^{(K-1)}$ plotted as functions of the prior mean μ. The intersection points of these three curves would determine stopping regions as a function of μ, reminiscent of those found in traditional sequential monitoring displays.

While the integration involved in computing $L_3^{(K-1)}$ is a bit awkward, it is not particularly difficult or high-dimensional. Rather, the primary hurdle arising from the backward induction process is the explosion in bookkeeping complexity arising from the resulting alternating sequence of integrations and minimizations. The number of Monte Carlo generations required to implement a backward induction solution also increases exponentially in the number of backward steps.

7.7.3 Forward sampling

The backward induction method's geometric explosion in both organizational and computational complexity as K increases motivates a search for an alternative approach. We now describe a forward sampling method which permits finding the optimal sequential decision in a prespecified class of such decisions. We implement this method when the class consists of decision rules of the following form: at each step k,

- for $E(\theta|y_1, \ldots, y_k) \le \gamma_{k,L}$, make decision d_1 (stop and choose the treatment);

- for $E(\theta|y_1,\ldots,y_k) > \gamma_{k,U}$, make decision d_2 (stop and choose placebo);
- for $\gamma_{k,L} < E(\theta|y_1,\ldots,y_k) \le \gamma_{k,U}$, make decision d_3 (continue sampling).

Carlin, Kadane, and Gelfand (1998) provide sufficient conditions to ensure that the optimal strategy is of this form; customary normal mean, binomial, and Poisson models meet these conditions.

Note that $\gamma_{k,L} \le \gamma_{k,U}$, with equality always holding for $k = K$ (i.e., $\gamma_{K,L} = \gamma_{K,U} \equiv \gamma_K$). This special structure motivates the following forward sampling algorithm, which for comparison with Section 7.7.2 is illustrated in the case $K = 2$. Suppose an independent sample $(\theta_j^*, y_{K-1,j}^*, y_{K,j}^*)$, $j = 1, \ldots, G$ is drawn via composition from the joint density $p(\theta, y_{K-1}, y_K) = p(\theta)f(y_{K-1}|\theta)f(y_K|\theta)$. When K = 2, decision rules can be indexed by the $2K$-1 = 5-dimensional vector $\boldsymbol{\gamma} = (\gamma_{K-2,L}, \gamma_{K-2,U}, \gamma_{K-1,L}, \gamma_{K-1,U}, \gamma_K)'$. Suppose that at the i^{th} iteration of the algorithm, a current estimate $\boldsymbol{\gamma}^{(i)}$ of the optimal rule of this form is available. For each of the simulated parameter-data values, the loss incurred under the current rule $\boldsymbol{\gamma}^{(i)}$ can be computed as follows:

$$
\begin{array}{rl}
\text{If } E(\theta) \le \gamma_{K-2,L}^{(i)}, & L_j = s_1^{(K-2)}(\theta_j^* - \theta_U) \\
\text{else if } E(\theta) > \gamma_{K-2,U}^{(i)}, & L_j = s_2^{(K-2)}(\theta_L - \theta_j^*) \\
\text{else if } E(\theta|y_{K-1,j}^*) \le \gamma_{K-1,L}^{(i)}, & L_j = s_1^{(K-1)}(\theta_j^* - \theta_U) + A_{K-1} \\
\text{else if } E(\theta|y_{K-1,j}^*) > \gamma_{K-1,U}^{(i)}, & L_j = s_2^{(K-1)}(\theta_L - \theta_j^*) + A_{K-1} \\
\text{else if } E(\theta|y_{K-1,j}^*, y_{K,j}^*) \le \gamma_K^{(i)}, & L_j = s_1^{(K)}(\theta_j^* - \theta_U) + A_{K-1} + A_K \\
\text{else,} & L_j = s_2^{(K)}(\theta_L - \theta_j^*) + A_{K-1} + A_K
\end{array}
\tag{7.19}
$$

Then $\bar{L} = \frac{1}{G}\sum_{j=1}^{G} L_j$ is a Monte Carlo estimate of the posterior expected loss under decision rule $\boldsymbol{\gamma}^{(i)}$. If the current estimate is not the minimum value, it is adjusted to $\boldsymbol{\gamma}^{(i+1)}$ and the above process is repeated. At convergence, the algorithm provides the desired optimal stopping boundaries.

Note that many of the $y_{K,j}^*$ values (and some of the $y_{K-1,j}^*$ values as well) will not actually figure in the determination of L, since they contribute only to the later conditions in the decision tree (7.19). These cases correspond to simulated trials that stopped before these later data values were actually observed; ignoring them in a given replication creates the proper marginal distribution of the parameter and the observed data (if any). Indeed, note that when $E(\theta) \notin (\gamma_{K-2,L}, \gamma_{K-2,U}]$ all of the sampled values fall into one of the first two cases in (7.19). Second, L is not likely to be smooth enough as a function of $\boldsymbol{\gamma}^{(i)}$ to allow its minimization via standard methods, such as a quasi-Newton method. Better results may be obtained by applying a more robust, univariate method (such as a golden section search) sequentially in each coordinate direction, i.e., within an iterated conditional modes algorithm (see Besag, 1986).

The main advantage of this approach over backward induction is clear: since the length of γ is $2k$ - 1, the computational burden in forward sampling grows *linearly* in K, instead of exponentially. The associated bookkeeping and programming is also far simpler. Finally, the forward sampling algorithm is easily adapted to more advanced loss structures. For example, switching back to the positive part loss functions (7.16). a change that would cause substantial complications in the backward induction solution, leads only to a corresponding change in the scoring rule (7.19).

A potential disadvantage of the forward sampling algorithm is that it replaces a series of simple minimizations over discrete finite spaces (a choice among 3 alternatives) with a single, more difficult minimization over a continuous space $(\Re^{2K-1})$. One might well wonder whether this minimization is any more feasible for large K than backward induction. Fortunately, even when the exact minimum is difficult to find, it should be possible to find a rule that is nearly optimal via grid search over γ-space, meaning that approximate solutions will be available via forward sampling even when backward induction is infeasible.

It is critical to get a good estimate of *each* of the components of the optimizing γ and not just the "components of interest" (typically, the early components of γ near the current time point), for otherwise there is a risk of finding a local rather than a global minimum. As such, one may want to sample not from the joint distribution $p(\theta, y_1, \ldots, y_K)$ but from some modification of it designed to allow more realizations of the latter stages of the decision tree (7.19). To correct for this, one must then make the appropriate importance sampling adjustment to the L_j scores in
By deliberately straying from the true joint distribution, one is sacrificing overall precision in the estimated expected loss in order to improve the estimate of the later elements of γ. As such, the phrase "unimportance sampling" might be used to refer to this process.

Discussion

In the case of a one-parameter exponential family likelihood, forward sampling may be used to find the optimal rule based solely on the posterior mean. In a very general parametric case, the optimal sequential strategy could be a complex function of the data, leading to a very large class of strategies to be searched. But what of useful intermediate cases - say, a normal likelihood with both the mean and the variance unknown? Here, the class of possibly optimal sequential rules includes arbitrary functions of the sufficient statistic vector $(\bar{y}, s^2)$. In such a case, for the particular loss function in question, it may be possible to reduce the class of possible optimal strategies. But for practical purposes it may not be necessary to find the optimal sequential strategy. Forward sampling can be used to evaluate general sequential strategies of any specified form. If the class of sequential strategies offered to the algorithm is reasonably broad, the best strategy in that class is likely to be quite good, even if it is not optimal in the class of all sequential strategies. This constitutes a distinct advantage of the forward sampling algorithm. When backward induction becomes infeasible, as it quickly does even in simple cases, it offers no advice about what sequential strategy to employ. Forward sampling, however, can be applied in such a situation with a restricted class of sequential strategies to find a good practical one even if it is not the best possible.

7.8 Spatial and spatio-temporal models

Statisticians are increasingly faced with the task of analyzing data that are geographically referenced, and often presented in the form of maps. Echoing a theme already common in this book, this increase has been driven by advances in computing; in this case, the advent of *geographic information systems* (GISs), sophisticated computer software packages that allow the simultaneous graphical display and summary of spatial data. But while GISs such as `ARC/INFO`, `ArcView`, and `MapInfo` are extraordinarily powerful for data manipulation and display, as of the present writing, they have little capacity for statistical modeling and inference.

To establish a framework for spatial modeling, we follow that given in the enormously popular textbook on the subject by Cressie (1993). Let $\mathbf{s} \in \Re^d$ be a location in d-dimensional space, and $X(\mathbf{s})$ be a (possibly vector-valued) data value observed at s. The full dataset can then be modeled as the multivariate random process

$$\{X(\mathbf{s}) : \mathbf{s} \in D\} , \tag{7.20}$$

where s varies over an index set $D \subset \Re^d$. Cressie (1993) then uses (7.20) to categorize spatial data into three cases.

- *geostatistical data,* where $X(s)$ is a random vector at a location s which varies *continuously* over D, a fixed subset of $\Re^d$ that contains a d-dimensional rectangle of positive volume;
- *lattice data,* where D is again a fixed subset (of regular or irregular shape), but now containing only countably many sites in $\Re^d$ normally supplemented by neighbor information;
- *point pattern data,* where now D is itself random; its index set gives the locations of random events that are the spatial point pattern. $X(s)$ itself can simply equal 1 for all $\mathbf{s} \in D$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a *marked point pattern process*).

An example of the first case might be a collection of measured oil or water reserves at various fixed *source* points $\mathbf{s}$, with the primary goal being to predict the reserve available at some unobserved *target* location t. In the second case, the observations might correspond to pixels on a computer screen, each of which has precisely four neighbors (up, down, left, and right), or perhaps average values of some variable over a geographic region (e.g., a county), each of whose neighbor sets contains those other counties that share a common border. Finally, point pattern data often arise as locations of certain kinds of trees in a forest, or residences of persons suffering from a particular disease (which in this case might be supplemented by age or other covariate information, producing a marked point pattern).

In the remainder of this section, we provide guidance on the Bayesian fitting of models for the first two cases above. While mining applications provided the impetus for much of the model development in spatial statistics, we find the term "geostatistical" somewhat misleading, since these methods apply much more broadly. Similarly, we dislike the term "lattice," since it connotes data corresponding to "corners" of a checkerboard-like grid. As such, we prefer to relabel these two cases more generically, according to the nature of the set D. Thus in Subsection 7.8.1 we consider models for geostatistical data, or what we call *point source data models,* where the locations are fixed points in space that may occur anywhere in D. Then in Subsection 7.8.2 we turn to lattice data models, or what we term *regional summary data models,* where the source data are counts or averages over geographical regions.

In the **S-plus** environment familiar to many statisticians, the maps library has for some time allowed display of point source data and regional summary data at the U.S. state or county level; more recently the **S+SpatialStats** add-on now enables easy access to many traditional geostatistical, lattice, and point pattern analysis tools (see Kaluzny et al., 1998). Our treatment is not nearly so comprehensive, but does suggest how Bayesian methods can pay dividends in a few of the most common spatial analysis settings.

7.8.1 Point source data models

We begin with some general notation and terminology. Let $X(\mathbf{s}_i)$ denote the spatial process associated with a particular variable (say, environmental quality), observed at locations $\{\mathbf{s}_i : \mathbf{s}_i \in D \subset \Re^d\}$. A basic starting point is to assume that $X(\mathbf{s}_i)$ is a *second-order stationary* process, i.e., that $E[X(\mathbf{s})] = \mu$ and $Cov[X(\mathbf{s}_i), X(\mathbf{s}_j)] \equiv C(\mathbf{s}_i - \mathbf{s}_j) < \infty$. C is called the *covariogram,* and is analogous to the autocovariance function in time series analysis. It then follows that $Var[X(\mathbf{s}_i) - X(\mathbf{s}_j)] \equiv 2\gamma(\mathbf{s}_i - \mathbf{s}_j)$ is a function of the separation vector $\mathbf{s}_i - \mathbf{s}_j$ alone, and is called the *variogram;* $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ itself is called the *semivariogram.* The variogram is said to be *isotropic* if it is a function solely of the Euclidean distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, d_{ij}; in this case $2\gamma(\mathbf{s}_i - \mathbf{s}_j) = 2\gamma(d_{ij})$, so that $C(\mathbf{s}_i - \mathbf{s}_j) = C(d_{ij})$.

We define the sill of the variogram as $\lim_{d\to\infty} 2\gamma(d)$, and the *nugget* as

$\lim_{d\to 0} 2\gamma(d)$. The latter need not be zero (i.e., the variogram may have a discontinuity at the origin) due to microscale variability or measurement error (Cressie, 1993, p.59). If $C(d) \to 0$ as $d \to \infty$, then $2\gamma(d) \to 2C(0)$, the sill. For a monotonic variogram that reaches its sill exactly, the distance at which this sill is reached is called the *range;* observations further apart than the range are uncorrelated. If the sill is reached only asymptotically, the distance such that $Corr[X(\mathbf{s}_i), X(\mathbf{s}_j)] = 0.05$ is called the *effective range.*

If we further assume a second-order stationary *Gaussian* process, our model can be written as

$$X(\mathbf{s}_i) = \mu + W(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \ , \tag{7.21}$$

where μ is again the grand mean, $W(s)$ accounts for spatial correlation, and $\epsilon(\mathbf{s})$ captures measurement error, $W(s)$ and $\epsilon(\mathbf{s})$ independent and normally distributed. To specify an isotropic spatial model, we might use a parametric covariance function, such as $Cov(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \sigma^2\rho(d_{ij};\phi)$. For example, we might choose $\rho(d_{ij};\phi) = \exp(-\phi d_{ij})$, $\phi > 0$, so that the covariance decreases with distance at constant exponential rate ϕ. The measurement error component is then often specified as $\epsilon(\mathbf{s}) \stackrel{iid}{\sim} N(0,\tau^2)$. From this model specification we may derive the variogram as

$$2\gamma(\tau^2,\sigma^2,\phi) = 2[\tau^2 + \sigma^2(1-\rho(d_{ij};\phi))] \ , \tag{7.22}$$

so that τ^2 is the nugget and $\tau^2 + \sigma^2$ is the sill.

Despite the occasionally complicated notation and terminology, equations (7.21) and (7.22) clarify that the preceding derivation determines only a four-parameter model. Bayesian model fitting thus proceeds once we specify the prior distribution for $\boldsymbol{\theta} = (\mu,\tau^2,\sigma^2,\phi)$ standard choices for these four components might be flat, inverse gamma, inverse gamma, and gamma, respectively (the latter since there is no hope for a conjugate specification for ϕ). Since this model is quite low-dimensional, any of the computational methods described in Chapter 5 might be appropriate, in particular, one could easily obtain draws $\{\boldsymbol{\theta}^{(g)}\ g = 1,\ldots G\}$ from $p(\boldsymbol{\theta}|\mathbf{X})$.

Bayesian kriging

While marginal posterior distributions such as $p(\phi|\mathbf{x})$ might be of substantial interest, historically the primary goal of such models has been to predict the unobserved value $X(\mathbf{t})$ at some target location $\mathbf{t}$. The traditional minimum mean squared error approach to this spatial prediction problem is called *kriging,* a rather curious name owing to the influential early work of South African mining engineer D.G. Krige (see e.g. Krige, 1951). Handcock and Stein (1993) and Diggle et al. (1998) present the Bayesian implementation of kriging, which is most naturally based on the predictive distribution of the targets given the sources. More specifically, suppose we seek predictions at target locations $(\mathbf{t}_1,\ldots,\mathbf{t}_m)$ given (spatially correlated) response data at source locations $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$. Following equations (7.21) and (7.22) above, the joint distribution of $\mathbf{X}_1 = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))'$ and $\mathbf{X}_2 = (X(\mathbf{t}_1), \ldots, X(\mathbf{t}_m))'$ is

$$[(\mathbf{X}_1, \mathbf{X}_2)' | \mu, \tau^2, \sigma^2, \phi] \sim MVN_{m+n}\left(\mu \mathbf{1}_{m+n}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) , \qquad (7.23)$$

where $\Sigma_{kl} \equiv Cov(\mathbf{X}_k, \mathbf{X}_l)$, k,1 = 1,2 based on our isotropic model (e.g., the ijth element of Σ_{11} is just $\sigma^2 \exp(-\phi d_{ij})$ where d_{ij} is the separation distance between $\mathbf{s}_i$ and $\mathbf{s}_j$ and similarly for the other blocks). Thus using standard multivariate normal distribution theory (see e.g. Guttman, 1982, pp.70-72), the conditional distribution of the target values $\mathbf{X}_2$ given the source values $\mathbf{X}_1$ and the model parameters is

$$[\mathbf{X}_2 | \mathbf{X}_1, \mu, \tau^2, \sigma^2, \phi] \sim MVN_m(\mu_{2.1}, \Sigma_{2.1}) .$$

where

$$\mu_{2.1} = \mu \mathbf{1}_m + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \mu \mathbf{1}_n), \text{ and } \Sigma_{2.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} .$$

Thus, marginalizing over the parameters, the predictive distribution we seek is given by

$$\begin{aligned} p(\mathbf{X}_2 | \mathbf{X}_1) &= \int p(\mathbf{X}_2 | \mathbf{X}_1, \mu, \tau^2, \sigma^2, \phi) p(\mu, \tau^2, \sigma^2, \phi | \mathbf{X}_1) d\mu d\tau^2 d\sigma^2 d\phi \\ &\approx \frac{1}{G} \sum_{g=1}^{G} p(\mathbf{X}_2 | \mathbf{X}_1, \mu^{(g)}, (\tau^2)^{(g)}, (\sigma^2)^{(g)}, \phi^{(g)}) , \qquad (7.24) \end{aligned}$$

where $(\mu^{(g)}, (\tau^2)^{(g)}, (\sigma^2)^{(g)}, \phi^{(g)})$ are the posterior draws obtained above. Point and interval estimates for the spatially interpolated values $X(\mathbf{t}_j)$ then arise from the appropriate marginal summaries of (7.24). Recall from Chapter 5 that we need not actually do the mixing in (7.24): merely drawing $\mathbf{X}_2^{(g)}$ values by composition from $p(\mathbf{X}_2 | \mathbf{X}_1, \mu^{(g)}, (\tau^2)^{(g)}, (\sigma^2)^{(g)}, \phi^{(g)})$ is sufficient to produce a histogram estimate of any component of $p(\mathbf{X}_2 | \mathbf{X}_1)$.

Anisotropy

If the spatial correlation between $X(\mathbf{s}_i)$ and $X(\mathbf{s}_j)$ depends not only on the length of the separation vector $\mathbf{h}_{ij} \equiv \mathbf{s}_i - \mathbf{s}_j$, but on its direction as well, the process $X(\mathbf{s})$ is said to be anisotropic. A commonly modeled form of anisotropy is geometric range anisotropy, where we assume there exists a positive definite matrix B such that the variogram takes the form

$$2\gamma(\mathbf{h}_{ij}) = 2\gamma((\mathbf{h}_{ij}' B \mathbf{h}_{ij})^{1/2}) .$$

Taking $B = \phi^2 I$ produces $(\mathbf{h}_{ij}' B \mathbf{h}_{ij})^{1/2} = \phi d_{ij}$ hence isotropy. Thus anisotropy replaces ϕ by the parameters in B; equation (7.22) generalizes to

$$2\gamma(\tau^2, \sigma^2, B) = 2[\tau^2 + \sigma^2(1 - \rho((\mathbf{h}_{ij}' B \mathbf{h}_{ij})^{1/2}))] , \qquad (7.25)$$

so that τ^2 and $\tau^2 + \sigma^2$ are again the nugget and sill, respectively.

In the most common spatial setting, $\Re^2$, the 2 x 2 symmetric matrix $B = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{pmatrix}$ completely defines the geometric anisotropy, and produces elliptical (instead of spherical) correlation contours in the plane having orientation ω defined implicitly by the relation

$$\cot(2\omega) = \frac{\beta_{11} - \beta_{22}}{2\beta_{12}} .$$

Also, the effective range in the direction η, r_η, is given by

$$r_\eta = \frac{c}{(\mathbf{h}'_\eta B \mathbf{h}_\eta)^{1/2}} ,$$

where $\mathbf{h}_\eta = (\cos\eta, \sin\eta)$, the unit vector in the direction of η and c is a constant such that $\rho(c) = 0.05$. Thus for our simple exponential ρ function above, we have $c = -\log(0.05) \approx 3$.

Ecker and Gelfand (1999) again recommend a simple product of independent priors for the components of $\boldsymbol{\theta} = (\mu, \tau^2, \sigma^2, B)$, suggesting a *Wishart(R, v)* prior for B with small degrees of freedom ν. Again, while the notation may look a bit frightening, the result in $\Re^2$ is still only a 6-parameter model (up from 4 in the case of isotropy). Estimation and Bayesian kriging may now proceed as before; see Ecker and Gelfand (1999) for additional guidance on pre-analytic exploratory plots, model fitting, and a real-data illustration. See also Handcock and Wallis (1994) for a spatio-temporal extension of Bayesian kriging, an application involving temperature shifts over time, and related discussion. Le, Sun, and Zidek (1997) consider Bayesian multivariate spatial interpolation in cases where certain data are missing by design.

7.8.2 Regional summary data models

In this subsection we generally follow the outline of Waller et al. (1997), and consider spatial data (say, on disease incidence) that are available as summary counts or rates for a defined region, such as a county, district, or census tract. Within a region, suppose that counts or rates are observed for subgroups of the population defined by socio-demographic variables such as gender, race, age, and ethnicity. Suppose further that for each subgroup within each region, counts or rates are collected within regular time intervals, e.g. annually, as in our example. Hence in general notation, an observation can be denoted by $y_{i\ell t}$. Here $i = 1, ..., I$ indexes the regions, $\ell = 1, ..., L$ indexes the subgroups, and $t = 1, ..., T$ indexes the time periods. In application, subgroups are defined through factor levels, so that if we used, say, three factors, subscript ℓ would be replaced by jkm.

We assume $y_{i\ell t}$ is a count arising from a probability model with a baseline rate and a region-, group-, and year-specific relative risk $\psi_{i\ell t}$. Based on a likelihood model for the vector of observed counts $\mathbf{y}$ given the vector of relative risks $\boldsymbol{\psi}$, and a prior model on the space of possible $\boldsymbol{\psi}$s, MCMC computational algorithms can estimate a posterior for $\boldsymbol{\psi}$ given $\mathbf{y}$, which is used to create a disease map. Typically, a map is constructed by color-coding the means, medians, or other summaries of the posterior distributions of the $\psi_{i\ell t}$.

Typically, the objective of the prior specification is to stabilize rate estimates by *smoothing* the crude map, though environmental equity concerns imply interest in *explaining* the rates. The crude map arises from the likelihood model alone using estimates (usually MLEs) of the $\psi_{i\ell t}$ based only on the respective $y_{i\ell t}$. Such crude maps often feature large outlying relative risks in sparsely populated regions, so that the map may be visually dominated by rates having very high uncertainty. Equally important, the maximum likelihood estimates fail to take advantage of the similarity of relative risks in adjacent or nearby regions. Hence, an appropriate prior for $\boldsymbol{\psi}$ will incorporate exchangeability and/or spatial assumptions which will enable the customary Bayesian "borrowing of strength" to achieve the desired smoothing.

The likelihood model assumes that, given $\boldsymbol{\psi}$, the counts $y_{i\ell t}$ are conditionally independent Poisson variables. The Poisson acts as an approximation to a binomial distribution, say, $y_{i\ell t} \sim Bin(n_{i\ell t}, p_{i\ell t})$ where $n_{i\ell t}$ is the known number of individuals at risk in county i within subgroup ℓ at time t, and $p_{i\ell t}$ is the associated disease rate. Often the counts are sufficiently large so that $y_{i\ell t}$ or perhaps $\sqrt{y_{i\ell t}}$ (or the Freeman-Tukey transform, $\sqrt{y_{i\ell t}} + \sqrt{y_{i\ell t}+1}$) can be assumed to approximately follow a normal density. In our case, however, the partitioning into subgroups results in many small values of $y_{i\ell t}$ (including several 0s), so we confine ourselves to the Poisson model, i.e. $y_{i\ell t} \sim Po(n_{i\ell t} p_{i\ell t})$.

We reparametrize to relative risk by writing $E_{i\ell t}\psi_{i\ell t} = n_{i\ell t}p_{i\ell t}$, where $E_{i\ell t}$ is the expected count in region i for subgroup ℓ at time t. Thus our basic model takes the form

$$y_{i\ell t} \sim Po(E_{i\ell t}\psi_{i\ell t}) \; .$$

In particular, if p^* is an overall disease rate, then $\psi_{i\ell t} = p_{i\ell t}/p^*$ and $E_{i\ell t} = n_{i\ell t}p^*$. The ψs are said to be *externally standardized* if p^* is obtained from another data source, such as a standard reference table (see e.g. Bernardinelli and Montomoli, 1992). The ψs are said to be *internally standardized* if p^* is obtained from the given dataset, e.g., if $p^* = \sum_{i\ell t} y_{i\ell t} / \sum_{i\ell t} n_{i\ell t}$. In our data example we lack an appropriate standard reference table, and therefore rely on the latter approach. Under external standardizing a product Poisson likelihood arises, while under internal standardizing the joint distribution of the $y_{i\ell t}$ is multinomial. However, since likelihood inference is unaffected by whether or not we condition on $\sum_{i\ell t} y_{i\ell t}$ (see e.g. Agresti, 1990, p. 455-56), it is common practice to retain the product Poisson likelihood.

The interesting aspect in the modeling involves the prior specification for $\psi_{i\ell t}$, or equivalently for $\mu_{i\ell t} \equiv \log \psi_{i\ell t}$. Models for $\mu_{i\ell t}$ can incorporate a variety of *main effect* and *interaction* terms. Consider the main effect for subgroup membership. In our example, with two levels for sex and two levels for race, it takes the form of a sex effect plus a race effect plus a sex-race interaction effect. In general, we can write this main effect as $\varepsilon_\ell = \mathbf{x}_\ell^T \boldsymbol{\beta}$ for an appropriate design vector $\mathbf{x}_\ell$ and parameter vector $\boldsymbol{\beta}$. In the absence of prior information we would take a vague, possibly flat, prior for $\boldsymbol{\beta}$. In many applications it would be sensible to assume additivity and write $\mu_{i\ell t} = \varepsilon_\ell + \varepsilon_{it}$. For instance, in our example this form expresses the belief that sex and race effects would not be affected by region and year. This is the only form we have investigated thus far though more complicated structures could, of course, be considered. Hence, it remains to elaborate ε_{it}.

Temporal modeling

As part of ε_{it}, consider a main effect for time δ_t. While we might specify δ_t as a parametric function (e.g., linear or quadratic), we prefer a qualitative form that lets the data reveal any presence of trend via the set of posterior densities for the δ_t. For instance, a plot of the posterior means or medians against time would be informative. But what sort of prior shall we assign to the set $\{\delta_t\}$? A simple choice used in Section 8.3 is again a flat prior. An alternative is an autoregressive prior with say AR(1) structure, that is, $\delta_t|\delta_{t-1} \sim N\left(\gamma_t + \rho(\delta_{t-1} - \gamma_{t-1})\,,\ \sigma_\delta^2\right)$. We might take $\gamma_t = \gamma t$, whence γ and σ_δ^2 are hyperparameters.

Spatial modeling

The main effect for regions, say η_i, offers many possibilities. For instance, we might have regional covariates collected in a vector $\mathbf{z}_i$ which contribute a component $h(\mathbf{z}_i)$. Typically h would be a specified parametric function $h(\mathbf{z}_i;\boldsymbol{\omega})$ possibly linear, i.e., $\mathbf{z}_i^T\boldsymbol{\omega}$. In any event, a flat prior for $\boldsymbol{\omega}$ would likely be assumed. In addition to or in the absence of covariate information we might include regional random effects θ_i to capture heterogeneity among the regions (Clayton and Bernardinelli, 1992; Bernardinelli and Montomoli, 1992). If so, since i arbitrarily indexes the regions, an exchangeable prior for the θ_is seems appropriate, i.e.,

$$\theta_i \overset{iid}{\sim} N\left(\kappa_\theta\ ,\ \frac{1}{\tau}\right)\ . \tag{7.26}$$

We then add a flat hyperprior for κ_θ, but require a proper hyperprior (typically a gamma distribution) for τ. Formally, with an improper prior such as

$\tau^{(d-3)/2}$ d = 0, 1, 2, the resultant joint posterior need not be proper (Hobert and Casella, 1996). Pragmatically, an informative prior for τ insures well behaved Markov chain Monte Carlo model fitting; see Subsection 8.3.3 for further discussion.

Viewed as geographic locations, the regions naturally suggest the possibility of spatial modeling (Clayton and Kaldor, 1987; Cressie and Chan, 1989; Besag, York and Mollié 1991; Clayton and Bernardinelli, 1992; Bernardinelli and Montomoli, 1992). As with exchangeability, such spatial modeling introduces association across regions. Of course we should consider whether, after adjusting for covariate effects, it is appropriate for geographic proximity to be a factor in the correlation among observed counts. If so, we denote the spatial effect for region i by ϕ_i. In the spirit of Clayton and Kaldor (1987) and Cressie and Chan (1989) who extend Besag (1974), we model the ϕ_i using a conditional autoregressive (CAR) model. That is, we assume the conditional density of $\phi_i|\phi_{j\neq i}$ is proportional to $\exp\left[-\frac{\lambda}{2}(a_i\phi_i - \sum_{j\neq i} w_{ij}\phi_j)^2\right]$, where $w_{ij} \geq 0$ is a weight reflecting the influence of ϕ_j on the expectation of ϕ_i, and $a_i > 0$ is a "sample size associated with region i. From Besag (1974) we may show that the joint density of the vector of spatial effects $\boldsymbol{\phi}$ is proportional to $\exp(-\frac{\lambda}{2}\boldsymbol{\phi}^T B\boldsymbol{\phi})$ where $B_{ii} = a_i$ and $B_{ij} = -a_i w_{ij}$. It is thus a proper multivariate normal density with mean 0 and covariance matrix B^{-1} provided B is symmetric and positive definite. Symmetry requires $a_i w_{ij} = a_j w_{ji}$. A proper gamma prior is typically assumed for λ.

Two special cases have been discussed in the literature. One, exemplified in Cressie and Chan (1989), Devine and Louis (1994), and Devine, Halloran and Louis (1994), requires a matrix of interregion distances d_{ij}. Then w_{ij} is set equal to $g(d_{ij})$ for a suitable decreasing function g. For example, Cressie and Chan (1989) choose g based on the estimated variogram of the observations. Since $w_{ij} = w_{ji}$, symmetry of B requires a_i constant. in fact, $a_i = g(0)$, which without loss of generality we can set equal to 1. A second approach, as in e.g. Besag, York, and Mollié (1991), defines a set ∂_i of *neighbors* of region i. Such neighbors can be defined as regions adjacent (contiguous) to region i, or perhaps as regions within a prescribed distance of region i. Let n_i be the number of neighbors of region i and let $w_{ij} = 1/n_i$ if $j \in \partial_i$, and 0 otherwise. Then, if $a_i = n_i$, B is symmetric. It is easy to establish that

$$\phi_i \mid \phi_{j\neq i} \sim N\left(\bar{\phi}_i\ ,\ \frac{1}{\lambda n_i}\right)\ , \tag{7.27}$$

where $\bar{\phi}_i = n_i^{-1}\sum_{j\in\partial_i}\phi_j$. It is also clear that B is singular since the sum of all the rows or of all the columns is 0. Thus, the joint density is improper. In fact, it is straightforward to show that $\boldsymbol{\phi}^T B\boldsymbol{\phi}$ can be written as a sum of pairwise differences of neighbors. Hence, the joint density is invariant to translation of the ϕ_is, again demonstrating its impropriety. The notions of interregion distance and neighbor sets can be combined as in Cressie and Chan (1989) by setting $w_{ij} = g(d_{ij})$ for $j \in \partial_i$. In the exercises below and our Section 8.3 case study, however, we confine ourselves to pairwise difference specification based upon adjacency, since this requires no inputs other than the regional map, and refer to such specifications as $CAR(\lambda)$ priors.

Typically the impropriety in the CAR prior is resolved by adding the sum-to-zero constraint $\sum_{i=1}^{I} \phi_i = 0$, which is awkward theoretically but simple to implement "on the fly" as part of an MCMC algorithm. That is, we simply subtract the current mean $\frac{1}{I}\sum_{i=1}^{I} \phi_i^{(g)}$ from all of the $\phi_i^{(g)}$ at the end of each iteration g. Again, adding a proper hyperprior for λ, proper posteriors for each of the ϕ_i ensue. The reader is referred to Eberly and Carlin (2000) for further insight into the relationship among identifiability, Bayesian learning, and MCMC convergence rates for spatial models using $CAR(\lambda$ priors.

Finally, note that we have chosen to model the spatial correlation in the prior specification using, in Besag's (1974) terminology, an auto-Gaussian model. One could envision introducing spatial correlation into the likelihood, i.e., directly amongst the $y_{i\ell t}$, resulting in Besag's auto-Poisson structure. However, Cressie (1993) points out that this model does *not* yield a Poisson structure marginally. Moreover, Cressie and Chan (1989) observe the intractability of this model working merely with y_is; for a set of $y_{i\ell t}$s, model fitting becomes infeasible.

A general form for the regional main effects is thus $\eta_i = h(\mathbf{z}_i; \boldsymbol{\omega}) + \theta_i + \phi_i$. This form obviously yields a likelihood in which θ_i and ϕ_i are not identifiable. However, the prior specification allows the possibility of separating these effects; the data can inform about each of them. A related issue is posterior integrability. One simple condition is that any effect in $\mu_{i\ell t}$ which is free of *i* and has a flat prior must have a fixed *baseline* level, such as 0. For example, $\mu_{i\ell t}$ cannot include an intercept. Equivalently, if both ϕ_i and θ_i are in the model then we must set $\kappa_\theta = 0$. Also, if δ_t terms are included we must include a constraint (say, $\delta_1 = 0$ or $\sum_t \delta_t = 0$), for otherwise we can add a constant to each ϕ_i and subtract it from each level of the foregoing effect without affecting the likelihood times prior, hence the posterior. It is also important to note that θ_i and ϕ_i may be viewed as surrogates for unobserved regional covariates. That is, with additional regional explanatory variables, they might not be needed in the model. From a different perspective, for models employing the general form for η_i, there may be strong collinearity between $\mathbf{z}_i$ and say ϕ_i, which again makes the ϕ_i difficult to identify, hence the models difficult to fit.

Spatio-temporal interaction

To this point, $\mu_{i\ell t}$ consists of the main effects ε_ℓ, δ_t, and η_i. We now turn to interactions, though as mentioned above the only ones we consider are spatio-temporal. In general, defining a space-time component is thorny because it is not clear how to reconcile the different scales. Bernardinelli, Clayton et al. (1995) assume the multiplicative form $(\nu_i^\theta + \nu_i^\phi)t$ for the heterogeneity by time and clustering by time effects. We again prefer a qualitative form and propose a *nested* definition, wherein heterogeneity effects and spatial effects are nested within time. In this way we can examine the evolution of spatial and heterogeneity patterns over time. We write these effects as $\theta_i^{(t)}$ and $\phi_i^{(t)}$. That is, $\theta_i^{(t)}$ is a random effect for the i^{th} region in year *t*. Conditional exchangeability given t would lead us to the prior $\theta_i^{(t)} \stackrel{iid}{\sim} N(\kappa_\theta^{(t)}, 1/\tau_t)$. Similarly, $\phi_i^{(t)}$ is viewed as a spatial effect for the i^{th} region in year t. Again we adopt a CAR prior but now given t as well. Hence, we assume $\phi_i^{(t)}|\phi_{j\neq i}^{(t)}$ has density proportional to $\exp\left[-\frac{\lambda_t}{2}(a_i\phi_i^{(t)} - \sum_{j\neq i} w_{ij}\phi_j^{(t)})^2\right]$. Again, proper gamma priors are assumed for the τ_t and the λ_t. Such definitions for $\theta_i^{(t)}$ and $\phi_i^{(t)}$ preclude the inclusion of main effects θ_i and ϕ_i in the model. Also, if $\phi_i^{(t)}$ appears in the model, then we cannot include δ_t terms that lack an informative prior, since for any fixed *t* the likelihood could not identify both. Similarly. if both $\theta_i^{(t)}$ and $\phi_i^{(t)}$ appear in the model, we must set $\kappa_\theta^{(t)} = 0$.

In summary, the most general model we envision for $\mu_{i\ell t}$ takes the form

$$\mu_{i\ell t} = \mathbf{x}_\ell^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\omega} + \theta_i^{(t)} + \phi_i^{(t)} \,. \tag{7.28}$$

In practice, (7.28) might well include more effects than we need. In our Section 8.3 case study, we investigate more parsimonious reduced models.

7.9 Exercises

1. Let $G_k(t)$ be defined as in Subsection 7.1.1. Show that the posterior mode of $G_k(t)$, denoted $\tilde{G}_k(t)$, has equal mass at exactly *k* mass points that are derived from an expression which for each *t* is the likelihood of independent, non-identically distributed Bernoulli variables.

2. Let G_k^* be a discrete distribution with equal mass at exactly *k* mass points, to be used as an estimate of G_k. Prove that if the Y_j are i.i.d., then, under integrated squared-error loss and subject to the elf of the ensemble producing G_k^*, the best estimates of the individual θ_j are the mass points themselves, with order assigned as for the observed Y_j. Also, show that the optimal ranks for the θ_j are produced by the ranked Y_j.

3. Compute and compare the mean and variance induced by edfs formed from the coordinate-specific MLEs, standard Bayes estimates, and func-

tion $\bar{G}_k$ for the normal/normal, Poisson/gamma, and beta/binomial two-stage models.

4. (Devroye, 1986, p.38) Suppose X is a random variable having cdf F, and Y is a truncated version of this random variable with support restricted to the interval $[a, b]$. Then Y has cdf

$$G(y) = \begin{cases} 0\,, & y < a \\ \frac{F(y)-F(a)}{F(b)-F(a)}\,, & a \le y \le b \\ 1\,, & y > b \end{cases}$$

Show that Y can be generated as $F^{-1}(F(a) + U[F(b) - F(a)])$, where U is a $Unif(0, 1)$ random variate. (This result enables "one-for-one" generation from truncated distributions for which we can compute F and F^{-1} either exactly or numerically.)

5. Consider the estimation of human mortality rates between ages x and $x + k$, where $x \geq 30$. Data available from a mortality study of a group of independent lives includes d_i, the number of deaths observed in the unit age interval $[x + i - 1,\ x + i]$, and e_i, the number of person-years the lives were under observation (*exposed*) in this interval. Thinking of the e_i as constants and under the simplifying assumption of a constant force of mortality θ_i over age interval i, one can show that

$$d_i | \theta_i \overset{iid}{\sim} Po(e_i \theta_i),\ \ i = 1, \ldots, k\ .$$

An estimate of θ_i is thus provided by $r_i = d_i / e_i$, the unrestricted maximum likelihood estimate of θ_i, more commonly known as the *raw* mortality rate. We wish to produce a *graduated* (smoothed) sequence of θ_is which conform to the *increasing* condition

$$\boldsymbol{\theta} \in S^{INC} = \{\boldsymbol{\theta} :\ 0 < \theta_1 < \cdots < \theta_k < B\}$$

for some fixed positive B.

Assume that the θ_is are an i.i.d. sample from a $G(\alpha, \beta)$ distribution *before* imposing the above constraints, where a is a known constant and β has an $IG(a, b)$ hyperprior distribution. Find the full conditionals for the θ_i and β needed to implement the Gibbs sampler in this problem. Also, describe how the result in the previous problem would be used in performing the required sampling.

6. Table 7.1 gives a dataset of male mortality experience originally presented and analyzed by Broffitt (1988). The rates are for one-year intervals, ages 35 to 64 inclusive (i.e., bracket i corresponds to age $34 + i$ at last birthday).

   (a) Use the results of the previous problem to obtain a sequence of graduated, strictly increasing mortality rate estimates. Set the upper bound $B = .025$ and use the data-based (empirical Bayes) hyperparameter

i	d_i	e_i	r_i	i	d_i	e_i	r_i
1	3	1771.5	0.0016935	16	4	1516.0	0.0026385
2	1	2126.5	0.0004703	17	7	1371.5	0.0051039
3	3	2743.5	0.0010935	18	4	1343.0	0.0029784
4	2	2766.0	0.0007231	19	4	1304.0	0.0030675
5	2	2463.0	0.0008120	20	11	1232.5	0.0089249
6	4	2368.0	0.0016892	21	11	1204.5	0.0091324
7	4	2310.0	0.0017316	22	13	1113.5	0.0116749
8	7	2306.5	0.0030349	23	12	1048.0	0.0114504
9	5	2059.5	0.0024278	24	12	1155.0	0.0103896
10	2	1917.0	0.0010433	25	19	1018.5	0.0186549
11	8	1931.0	0.0041429	26	12	945.0	0.0126984
12	13	1746.5	0.0074435	27	16	853.0	0.0187573
13	8	1580.0	0.0050633	28	12	750.0	0.0160000
14	2	1580.0	0.0012658	29	6	693.0	0.0086580
15	7	1467.5	0.0047700	30	10	594.0	0.0168350

Table 7.1 *Raw mortality data (Broffitt, 1988)*

values $\alpha = \hat{\alpha} \equiv \bar{r}^2/(s_r^2 - \bar{r}\sum_{i=1}^k e_i^{-1}/k)$, $a = 3.0$, and $b = \hat{\alpha}/(2\bar{r})$, where $\bar{r} = \sum_{i=1}^k r_i/k$ and $s_r^2 = \sum_{i=1}^k (r_i - \bar{r})^2/(k-1)$. These values of a and b correspond to a rather vague hyperprior having mean and standard deviation both equal to $\hat{\beta} = \bar{r}/\hat{\alpha}$. (In case you've forgotten, $\hat{\alpha}$ and $\hat{\beta}$ are the method of moments EB estimates derived in Chapter 3, Problem 9.)

(b) Do the graduated rates you obtained form a sequence that is not only increasing, but *convex* as well? What modifications to the model would guarantee such an outcome?

7. *Warning:* In this problem, we use the **BUGS** parametrizations of the normal and gamma distributions: namely, a $N(\mu, \tau)$ has *precision* (not variance) τ, and a $G(\alpha, \beta)$ has mean α/β and variance α/β^2.

   Consider the data in Table 7.2, also available on the web in **WinBUGS** format at **http://www.biostat.umn.edu/~brad/data-html**. Originally presented by Wakefield et al. (1994), these data record the plasma concentration Y_{ij} of the drug Cadralazine at various time lags x_{ij} following the administration of a single dose of 30 mg in 10 cardiac failure patients. Here, $i = 1, \ldots, 10$ indexes the patient, while $j = 1, \ldots, n_i$ indexes the observations, $5 \leq n_i \leq 8$. These authors suggest a "one-compartment" nonlinear pharmacokinetic model wherein the mean plasma concentra-

patient	no. of hours following drug administration 2	4	6	8	10	24	28	32
1	1.09	0.75	0.53	0.34	0.23	0.02	–	–
2	2.03	1.28	1.20	1.02	0.83	0.28	–	–
3	1.44	1.30	0.95	0.68	0.52	0.06	–	–
4	1.55	0.96	0.80	0.62	0.46	0.08	–	–
5	1.35	0.78	0.50	0.33	0.18	0.02	–	–
6	1.08	0.59	0.37	0.23	0.17	–	–	–
7	1.32	0.74	0.46	0.28	0.27	0.03	0.02	–
8	1.63	1.01	0.73	0.55	0.41	0.01	0.06	0.02
9	1.26	0.73	0.40	0.30	0.21	–	–	–
10	1.30	0.70	0.40	0.25	0.14	–	–	–

Table 7.2 *Cadralazine concentration data (Wakefield et al., 1994)*

tion $\eta_{ij}(x_{ij})$ at time x_{ij} is given by

$$\eta_{ij}(x_{ij}) = 30\alpha_i^{-1}\exp(-\beta_i x_{ij}/\alpha_i) \ .$$

Subsequent unpublished work by these same authors suggests this model is best fit on the log scale. That is, we suppose

$$Z_{ij} \equiv \log Y_{ij} = \log \eta_{ij}(x_{ij}) + \epsilon_{ij} \ ,$$

where $\epsilon_{ij} \overset{ind}{\sim} N(0, \tau_i)$. The mean structure for the Z_{ij}'s thus emerges as

$$\begin{aligned} \log \eta_{ij}(x_{ij}) &= \log\left[30\alpha_i^{-1}\exp(-\beta_i x_{ij}/\alpha_i)\right] \\ &= \log 30 - \log \alpha_i - \beta_i x_{ij}/\alpha_i \\ &= \log 30 - a_i - \exp(b_i - a_i)x_{ij} \ , \end{aligned}$$

where $a_i = \log \alpha_i$ and $b_i = \log \beta_i$.

(a) Assuming the two sets of random effects are independently distributed as $a_i \overset{iid}{\sim} N(\mu_a, \tau_a)$ and $b_i \overset{iid}{\sim} N(\mu_b, \tau_b)$, use **BUGS** or **WinBUGS** to analyze these data (see Subsection C.3.3 in Appendix C, or go directly to the program's webpage, http://www.mrc-bsu.cam.ac.uk/bugs/). Adopt a vague prior structure, such as independent G(0.0001, 0.0001) priors for the τ_i, independent N(0, 0.0001) priors for μ_a and μ_b, and G(1, 0.04) priors for τ_a and τ_b (the latter as encouraged the original authors).

Note that the full conditional distributions of the random effects are not simple conjugate forms nor guaranteed to be log-concave, so **BUGS'** Metropolis capability is required. Besides the usual posterior summaries and convergence checks, investigate the acceptance rate of the

i	x_{1i}	x_{2i}	y_i	i	x_{1i}	x_{2i}	y_i
1	3.7	0.825	1	21	0.4	2	0
2	3.5	1.09	1	22	0.95	1.36	0
3	1.25	2.5	1	23	1.35	1.35	0
4	0.75	1.5	1	24	1.5	1.36	0
5	0.8	3.2	1	25	1.6	1.78	1
6	0.7	3.5	1	26	0.6	1.5	0
7	0.6	0.75	0	27	1.8	1.5	1
8	1.1	1.7	0	28	0.95	1.9	0
9	0.9	0.75	0	29	1.9	0.95	1
10	0.9	0.45	0	30	1.6	0.4	0
11	0.8	0.57	0	31	2.7	0.75	1
12	0.55	2.75	0	32	2.35	0.03	0
13	0.6	3	0	33	1.1	1.83	0
14	1.4	2.33	1	34	1.1	2.2	1
15	0.75	3.75	1	35	1.2	2	1
16	2.3	1.64	1	36	0.8	3.33	1
17	3.2	1.6	1	37	0.95	1.9	0
18	0.85	1.415	1	38	0.75	1.9	0
19	1.7	1.06	0	39	1.3	1.625	1
20	1.8	1.8	1				

Table 7.3 *Finney's vasoconstriction data*

Metropolis algorithm, and the predictive distribution of $Y_{4,7}$, the missing observation on patient 4 at 28 hours after drug administration. Are any patients "outliers" in some sense?

(b) The assumption that the random effects a_i and b_i are independent within individuals was probably unrealistic. Instead, follow the original analysis of Wakefield et al. (1994) and assume the $\boldsymbol{\theta}_i \equiv (a_i, b_i)'$ are i.i.d. from a $N_2(\boldsymbol{\mu}, \Omega)$ distribution, where $\boldsymbol{\mu} = (\mu_a, \mu_b)$. Again adopt a corresponding vague conjugate prior specification, namely $\boldsymbol{\mu} \sim N_2(0, C)$ and $\Omega \sim \text{Wishart}((\rho R)^{-1}, \rho)$ with $C \approx 0$, $\rho = 2$, and $R = \text{Diag}(0.04,\ 0.04)$.

Describe the changes (if any) in your answers from those in part (a).

Summarize your experience with **WinBUGS**. Did it perform as indicated in the documentation? What changes would you suggest to improve its quality, generality, and ease of use in applied data analysis settings?

8. Suppose we have a collection of binary responses $y_i \in \{0, 1\}$, $i = 1, \ldots, n$, and associated k-dimensional predictor variables $\mathbf{x}_i$. Define the latent

variables y_i^* as

$$y_i^* = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \; i = 1, \ldots, n ,$$

where the ϵ_i are independent mean-zero errors having cumulative distribution function F, and $\boldsymbol{\beta}$ is a k-dimensional regression parameter. Consider the model

$$Y_i = \begin{cases} 0, & \text{if } Y_i^* \leq 0 \\ 1, & \text{if } Y_i^* > 0 \end{cases} .$$

(a) Show that if F is the standard normal distribution, this model is equivalent to the usual *probit* model for $p_i = P(Y_i = 1)$.

(b) Under a $N(\boldsymbol{\mu}, \Sigma)$ prior for $\boldsymbol{\beta}$, find the full conditional distributions for $\boldsymbol{\beta}$ and the Y_i^*, $i = 1, \ldots, n$.

(c) Use the results of the previous part to find the posterior distribution for $\boldsymbol{\beta}$ under a flat prior for the data in Table 7.3. Originally given by Finney (1947), these data give the presence or absence of vasoconstriction in the skin of the fingers following inhalation of a certain volume of air, x_1, at a certain average rate, x_2. (Include an intercept in your model, so $k = 3$.)

(d) How would you modify your computational approach if, instead of the probit model, we wished to fit the *logit* (logistic regression) model?

(e) Perform the actual MCMC fitting of the logistic and probit models using the **BUGS** or **WinBUGS** language. Do the fits differ significantly? (*Hint:* Follow the approach of Chapter 5, problem 13.)

9. Suppose that in the context of the previous problem, the data y_i are not merely binary but *ordinal, i.e.,* $y_i \in \{0, 1 \ldots. m\}$, $i = 1, \ldots, n$. We extend our previous model to

$$Y_i = \begin{cases} 0, & \text{if } Y_i^* \leq \gamma_1 \\ 1, & \text{if } \gamma_1 < Y_i^* \leq \gamma_2 \\ \vdots & \vdots \\ m-1, & \text{if } \gamma_{m-1} < Y_i^* \leq \gamma_m \\ m, & \text{if } Y_i^* > \gamma_m \end{cases} .$$

(a) Suppose we wish to specify G as the cdf of Y_i (e.g., $G \sim Bin(m, p_i)$). Find an expression for the appropriate choice of "bin boundaries" γ_j $j = 1, \ldots m$, and derive the new full conditional distributions for $\boldsymbol{\beta}$ and the Y_i^* in this case.

(b) Now suppose instead that we wish to leave G unspecified. Can we simply add all of the γ_js into the sampling order and proceed? What are the appropriate full conditional distributions to use for the γ_j?

10. The data in Table 7.4 were originally analyzed by Tolbert et al. (2000), and record the horizontal s_1 and vertical s_2 coordinates (in meters

station i	latitude s_{1i}	longitude s_{2i}	observed data $X(\mathbf{s}_i)$
1	−18.9805	359.3649	0.085
2	−16.9316	359.1380	0.069
3	−15.1309	357.1999	0.084
4	−12.5863	356.9081	0.104
5	−12.0162	356.5347	0.111
6	−9.9322	355.4129	0.069
7	−9.7132	364.1625	0.049
8	−13.1500	350.8925	0.053
9	−11.2352	358.3001	0.082
10	−9.8524	359.5717	0.065
A	−13.1383	356.5103	–
B	−13.9711	356.0425	–

Table 7.4 *Ozone data at ten monitoring stations, Atlanta metro area, 6/3/95*

$\times 10^4$) of ten ozone monitoring stations in the greater Atlanta metro area. Figure 7.4 on page 274 shows the station locations on a greater Atlanta zip code map. The table also gives the ten corresponding ozone measurements $X(\mathbf{s}_i)$ (one-hour max) for June 3, 1995.

(a) Using the isotropic Bayesian kriging model of Section 7.8.1 applied to $Z(\mathbf{s}_i) = \log X(\mathbf{s}_i)$, find the posterior distribution of the spatial smoothing parameter ϕ and the true ozone concentrations at target locations A and B, whose coordinates are also given in Table 7.4. (As can be seen from Figure 7.4, these are two points on opposite sides of the same Atlanta city zip code.) Set $\tau^2 = 0$, and use a flat prior on μ, a vague IG(3, 0.5) prior (mean 1, variance 1) on σ^2, and a vague G(.001,100) (mean .1, variance 10) prior on ϕ. Are your results consistent with the mapped source data?

(b) Why might an anisotropic model be more appropriate here? Fit this model, using a vague Wishart prior for the matrix B, and compare its fit to the isotropic model in part (a).

11. Table 7.5 presents observed (Y_i) and expected (E_i) cases of lip cancer in 56 counties in Scotland, which have been previously analyzed by Clayton and Kaldor (1987) and Breslow and Clayton (1993). The table also shows a spatial covariate, x_i, the percentage of each county's population engaged in agriculture, fishing, or forestry (AFF), the observed standardized mortality ratio (SMR) for the county, $100Y_i/E_i$, and the

i	Y_i	E_i	x_i	SMR	Adjacent counties (∂_i)
1	9	1.4	16	652.2	5, 9, 11, 19
2	39	8.7	16	450.3	7, 10
3	11	3.0	10	361.8	6, 12
4	9	2.5	24	355.7	18, 20, 28
5	15	4.3	10	352.1	1, 11, 12, 13, 19
6	8	2.4	24	333.3	3, 8
7	26	8.1	10	320.6	2, 10, 13, 16, 17
8	7	2.3	7	304.3	6
9	6	2.0	7	303.0	1, 11, 17, 19, 23, 29
10	20	6.6	16	301.7	2, 7, 16, 22
11	13	4.4	7	295.5	1, 5, 9, 12
12	5	1.8	16	279.3	3, 5, 11
13	3	1.1	10	277.8	5, 7, 17, 19
14	8	3.3	24	241.7	31, 32, 35
15	17	7.8	7	216.8	25, 29, 50
16	9	4.6	16	197.8	7, 10, 17, 21, 22, 29
17	2	1.1	10	186.9	7, 9, 13, 16, 19, 29
18	7	4.2	7	167.5	4, 20, 28, 33, 55, 56
19	9	5.5	7	162.7	1, 5, 9, 13, 17
20	7	4.4	10	157.7	4, 18, 55
21	16	10.5	7	153.0	16, 29, 50
22	31	22.7	16	136.7	10, 16
23	11	8.8	10	125.4	9, 29, 34, 36, 37, 39
24	7	5.6	7	124.6	27, 30, 31, 44, 47, 48, 55, 56
25	19	15.5	1	122.8	15, 26, 29
26	15	12.5	1	120.1	25, 29, 42, 43
27	7	6.0	7	115.9	24, 31, 32, 55
28	10	9.0	7	111.6	4, 18, 33, 45

Table *7.5 Observed SMRs for Lip Cancer in 56 Scottish Counties*

the spatial relation of each county to the others via the indices of those counties adjacent to it. The raw data and the AFF covariate are mapped in Figure 7.5 on page 274.

(a) Following Section 7.8.2, fit the model $Y_i \sim Po(E_i e^{\mu_i})$ where

$$\mu_i = \alpha x_i/10 + \theta_i$$

and the θ_i have the exchangeable prior (7.26). Estimate the posterior distributions of κ_θ and α, and comment on your findings.

i	Y_i	E_i	x_i	SMR	Adjacent counties (∂_i)
29	16	14.4	10	111.3	9, 15, 16, 17, 21, 23, 25, 26, 34, 43, 50
30	11	10.2	10	107.8	24, 38, 42, 44, 45, 56
31	5	4.8	7	105.3	14, 24, 27, 32, 35, 46, 47
32	3	2.9	24	104.2	14, 27, 31, 35
33	7	7.0	10	99.6	18, 28, 45, 56
34	8	8.5	7	93.8	23, 29, 39, 40, 42, 43, 51, 52, 54
35	11	12.3	7	89.3	14, 31, 32, 37, 46
36	9	10.1	0	89.1	23, 37, 39, 41
37	11	12.7	10	86.8	23, 35, 36, 41, 46
38	8	9.4	1	85.6	30, 42, 44, 49, 51, 54
39	6	7.2	16	83.3	23, 34, 36, 40, 41
40	4	5.3	0	75.9	34, 39, 41, 49, 52
41	10	18.8	1	53.3	36, 37, 39, 40, 46, 49, 53
42	8	15.8	16	50.7	26, 30, 34, 38, 43, 51
43	2	4.3	16	46.3	26, 29, 34, 42
44	6	14.6	0	41.0	24, 30, 38, 48, 49
45	19	50.7	1	37.5	28, 30, 33, 56
46	3	8.2	7	36.6	31, 35, 37, 41, 47, 53
47	2	5.6	1	35.8	24, 31, 46, 48, 49, 53
48	3	9.3	1	32.1	24, 44, 47, 49
49	28	88.7	0	31.6	38, 40, 41, 44, 47, 48, 52, 53, 54
50	6	19.6	1	30.6	15, 21, 29
51	1	3.4	1	29.1	34, 38, 42, 54
52	1	3.6	0	27.6	34, 40, 49, 54
53	1	5.7	1	17.4	41, 46, 47, 49
54	1	7.0	1	14.2	34, 38, 49, 51, 52
55	0	4.2	16	.0	18, 20, 24, 27, 56
56	0	1.8	10	.0	18, 24, 30, 33, 45, 55

Table 7.5 *(continued)*

(b) Repeat the previous analysis using the mean structure

$$\mu_i = \alpha x_i/10 + \phi_i$$

where the ϕ_i have the CAR adjacency prior (7.27). How do your results change?

(c) Conduct an informal comparison of these two models using the DIC criterion (6.25); which model emerges as better? Also use equation (6.24) to compare the two effective model sizes p_D.

(d) How might we determine whether significant spatial structure remains in the data after accounting for the covariate?

(e) Suppose we wanted to fit a model with *both* the heterogeneity terms (θ_i) and the clustering terms (ϕ_i). What constraints or modifications to the prior distributions would be required to insure model identifiability (and hence, MCMC sampler convergence)?

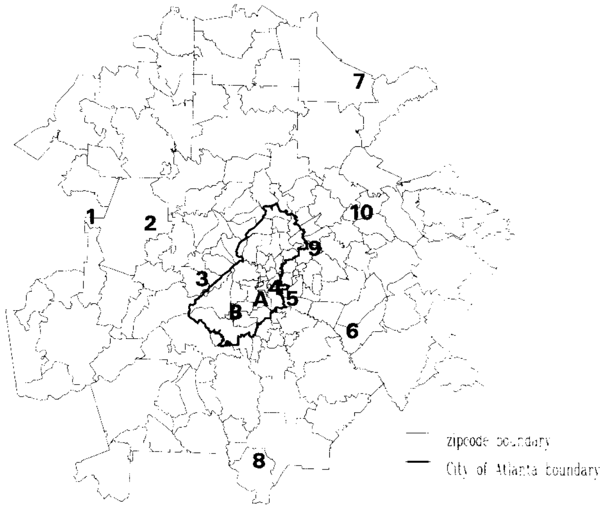


Figure 7.4 Ten ozone monitoring stations (source locations) and two target locations, greater Atlanta.

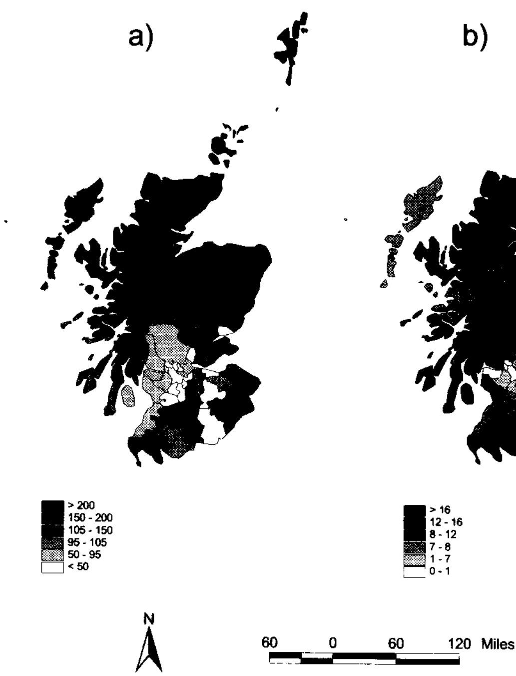


Figure 7.5 *Scotland lip cancer data: a) crude standardized mortality ratios (observed/expected* × *100); b) AFF covariate values.*

CHAPTER 8

Case studies

Most skills are best learned not through reading or quiet reflection, but through actual practice. This is certainly true of data analysis, an area with few hard-and-fast rules that is as much art as science. It is especially true of Hayes and empirical Bayes data analysis, with its substantial computing component and the challenging model options encouraged by the breadth of the methodology. The recent books edited by by Gatsonis et al. (1993, 1995, 1997, 1999, 2000), Gilks et al. (1996), and Berry and Stangl (1996) offer a broad range of challenging and interesting case studies in Bayesian data analysis, most using modern MCMC computational methods.

In this chapter we provide three fully worked case studies arising from real-world problems and datasets. While all have a distinctly biomedical flavor, they cover a broad range of methods and data types that transfer over to applications in agriculture, economics, social science, and industry. All feature data with complex dependence structures (often introduced by an ordering in time, space, or both), which complicate classical analysis and suggest a hierarchical Bayes or EB approach as a more viable alternative.

Our first case study evaluates drug-induced response in a series of repeated measurements on a group of AIDS patients using both longitudinal modeling and probit regression. We then compare the survival times of patients in each drug and response group. The second study involves sequential analysis, in the context of monitoring a clinical trial. The results of using carefully elicited priors motivates a comparison with those obtained with vague priors, as well as the sort of robust analysis suggested in Subsection 6.1.2. Finally, our third case study concerns the emerging field of spatial statistics, and illustrates the predictive approach to model selection described in Subsection 6.5. Here we obtain estimates and predictions for county-specific lung cancer rates by gender and race over a 21-year period in the U.S. state of Ohio. We also assess the impact of a potential air contamination source on the rates in the areas that surround it.

8.1 Analysis of longitudinal AIDS data

8.1.1 Introduction and background

Large-scale clinical trials often require extended follow-up periods to evaluate the clinical efficacy of a new treatment, as measured for example by survival time. The resulting time lags in reporting the results of the trial to the clinical and patient communities are especially problematic in AIDS research, where the life-threatening nature of the disease heightens the need for rapid dissemination of knowledge. A common approach for dealing with this problem is to select an easily-measured biological marker, known to be predictive of the clinical outcome, as a surrogate endpoint. For example, the number of CD4 lymphocytes per cubic millimeter of drawn blood has been used extensively as a surrogate marker for progression to AIDS and death in drug efficacy studies of HIV-infected persons. Indeed, an increase in the average CD4 count of patients receiving one of the drugs in our study (didanosine) in a clinical trial was the principal basis for its limited licensing in the United States.

Several investigators, however, have questioned the appropriateness of using CD4 count in this capacity. For example, Lin et al. (1993) showed that CD4 count was not an adequate surrogate for first opportunistic infection in data from the Burroughs Wellcome 02 trial, nor was it adequate for the development of AIDS or death in AIDS Clinical Trials Group (ACTG) Protocol 016 data. Choi et al. (1993) found CD4 count to be an incomplete surrogate marker for AIDS using results from ACTG Protocol 019 comparing progression to AIDS for two dose groups of zidovudine (AZT) and placebo in asymptomatic patients. Most recently, the results of the joint Anglo-French "Concorde" trial (Concorde Coordinating Committee, 1994) showed that, while asymptomatic patients who began taking AZT immediately upon randomization did have significantly higher CD4 counts than those who deferred AZT therapy until the onset of AIDS-related complex (ARC) or AIDS itself, the survival patterns in the two groups were virtually identical.

We analyze data from a trial involving 467 persons with advanced HIV infection who were randomly assigned to treatment with the antiretroviral drugs didanosine (ddI) or zalcitabine (ddC). This study, funded by the National Institutes of Health (NIH) and conducted by the Community Programs for Clinical Research on AIDS (CPCRA), was a direct randomized comparison of these two drugs in patients intolerant to, or failures on, AZT. The details of the conduct of the ddI/ddC study are described elsewhere (Abrams et al., 1994; Goldman et al., 1996); only the main relevant points are given here. The trial enrolled HIV-infected patients with AIDS or two CD4 counts of 300 or less, and who fulfilled specific criteria for AZT intolerance or failure. CD4 counts were recorded at study entry and again at the 2, 6, 12, and 18 month visits (though some of these observations are missing for many individuals). The study measured several outcome variables; we consider only total survival time.

Our main goal is to analyze the association among CD4 count, survival time, drug group, and AIDS diagnosis at study entry (an indicator of disease progression status). Our population of late-stage patients is ideal for this purpose, since it contains substantial information on both the actual (survival) and surrogate (CD4 count) endpoints.

8.1.2 Modeling of longitudinal CD4 counts

Let Y_{ij} denote the j^{th} CD4 measurement on the i^{th} individual in the study, $j = 1, \ldots, s_i$ and $i = 1, \ldots, n$. Arranging each individual's collection of observations in a vector $Y_i = (Y_{i1}, \ldots, Y_{is_i})^T$ we attempt to fit the mixed effects model (7.10), namely,

$$Y_i = \mathbf{X}_i \alpha + \mathbf{W}_i \beta_i + \epsilon_i \ ,$$

where as in Section 7.4, $\mathbf{X}_i$ is an $s_i \times p$ design matrix, α is a $p \times 1$ vector of fixed effects, $\mathbf{W}_i$ is an $s_i \times q$ design matrix (q typically less than p), and β_i is a $q \times 1$ vector of subject-specific random effects, usually assumed to be normally distributed with mean vector 0 and covariance matrix V. We assume that given β_i, the components of ϵ_i are independent and normally distributed with mean 0 and variance σ^2 (marginalizing over β_i, the Y_i components are again correlated, as desired). The β_is capture subject-specific effects, as well as introduce additional variability into the model.

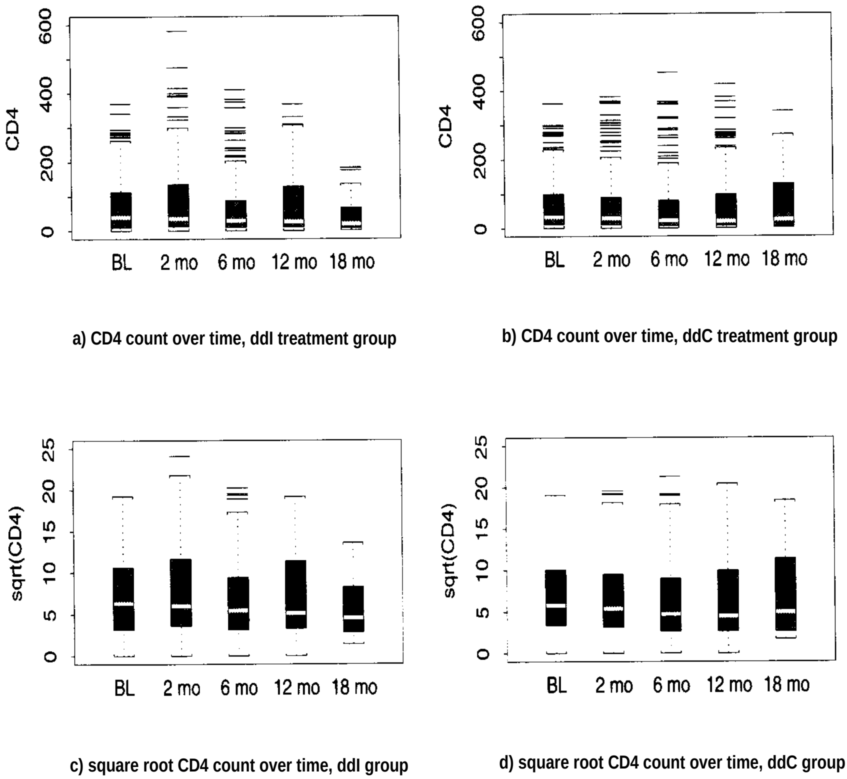
Since we wish to detect a possible increase in CD4 count two months after baseline, we fit a model that is linear but with possibly different slopes before and after this time. Thus the subject-specific design matrix $\mathbf{W}_i$ for patient i in equation (7.10) has j^{th} row

$$\mathbf{w}_{ij} = (1 \ , \ t_{ij} \ , \ (t_{ij} - 2)^+) \ ,$$

where $t_{ij} \in \{0, 2, 6, 12, 18\}$ and $z^+ = \max(z, 0)$. Hence the three columns of $\mathbf{W}_i$ correspond to individual-level intercept, slope, and change in slope following the changepoint, respectively. We account for the effect of covariates by including them in the fixed effect design matrix $\mathbf{X}_i$. Specifically, we set

$$\mathbf{X}_i = (\mathbf{W}_i \mid d_i \mathbf{W}_i \mid a_i \mathbf{W}_i) \ , \tag{8.1}$$

where d_i is a binary variable indicating whether patient i received ddI ($d_i = 1$) or ddC($d_i = 0$), and a_i is another binary variable telling whether the patient was diagnosed as having AIDS at baseline ($a_i = 1$) or not ($a_i = 0$). Notice from equation (8.1) that we have $p = 3q = 9$; the two covariates are being allowed to affect any or all of the intercept, slope, and change in slope of the overall population model. The corresponding elements of the α vector then quantify the effect of the covariate on the form



a) CD4 count over time, ddI treatment group

b) CD4 count over time, ddC treatment group

c) square root CD4 count over time, ddI group

d) square root CD4 count over time, ddC group

Figure 8.1 *Exploratory plots of CD4 count, ddI/ddC data.*

of the CD4 curve. In particular, our interest focuses on the α parameters corresponding to drug status, and whether they differ from 0.

Adopting the usual exchangeable normal model for the random effects, we obtain a likelihood of the form

$$\prod_{i=1}^{n} N_{s_i}(Y_i | \mathbf{X}_i\alpha + \mathbf{W}_i\beta_i, \sigma^2\mathbf{I}_{s_i}) \prod_{i=1}^{n} N_3(\beta_i | \mathbf{0}, \mathbf{V}) , \qquad (8.2)$$

where $N_k(\cdot|\mu, \boldsymbol{\Sigma})$ denotes the k-dimensional normal distribution with mean vector μ and covariance matrix $\boldsymbol{\Sigma}$. To complete our Bayesian model specification, we adopt independent prior distributions $N_9(\alpha|c, \mathbf{D})$, $IG(\sigma^2|a, b)$, and $IW(\mathbf{V}|(\rho\mathbf{R})^{-1}, \rho)$, where IG and IW denote the inverse gamma and inverse Wishart distributions, respectively (see Appendix A for details).

Turning to the observed data in our study, boxplots of the individual CD4 counts for the two drug groups, shown in Figures 8.1(a) and (b), indicate a high degree of skewness toward high CD4 values. This, combined with the count nature of the data, suggests a square root transformation for each group. As Figures 8.1(c) and (d) show, this transformation improves matters considerably. The sample medians, shown as white horizontal bars on the boxplots, offer reasonable support for our assumption of a linear decline in square root CD4 after two months. The sample sizes at the five time points, namely (230. 182, 153, 102, 22) and (236, 186, 157, 123, 14) for the ddI and ddC groups. respectively, indicate an increasing degree of missing data. The total number of observations is 1405, an average of approximately 3 per study participant.

Prior determination

Values for the parameters of our priors (the hyperparameters) may sometimes be determined from past datasets (Lange, Carlin, and Gelfand, 1992), or elicited from clinical experts (Chaloner et al., 1993; see also Section 8.2). In our case. since little is reliably known concerning the CD4 trajectories of late-stage patients receiving ddI or ddC, we prefer to choose hyperparameter values that lead to fairly vague, minimally informative priors. Care must be taken, however, to ensure that this does not lead to an improper posterior distribution. For example, taking p extremely small and D extremely large would lead to confounding between the fixed and random effects. (To see this, consider the simple but analogous case of the oneway layout, where $Y_i \stackrel{iid}{\sim} N(\mu + \alpha_i, \sigma^2)$ i = 1, . . ., n, and we place noninformative priors on *both* μ and the α_i. To avoid this while still allowing the random effects a reasonable amount of freedom, we adopt the rule of thumb wherein $\rho = n/20$ and

$$\mathbf{R} = Diag((r_1/8)^2, (r_2/8)^2, (r_3/8)^2) \ ,$$

where r_i is the total range of plausible parameter values across the individuals. Since $\mathbf{R}$ is roughly the prior mean of $\mathbf{V}$, this gives a ± 2 prior standard deviation range for β_i that is half of that plausible in the data. Since in our case Y_{ij} corresponds to square root CD4 count in late-stage HIV patients, we take $r_1 = 16$ and $r_2 = r_3 = 2$, and hence our rule produces $p = 24$ and $\mathbf{R} = Diag(2^2, (.25)^2, (.25)^2)$.

Turning to the prior on σ^2 we take $a = 3$ and $b = .005$, so that σ^2 has both mean and standard deviation equal to 10^2, a reasonably vague (but still proper) specification. Finally, for the prior on α, we set

$$\begin{aligned} c &= (10, 0, 0, 0, 0, 0, -3, 0, 0) \ , \quad \text{and} \\ \mathbf{D} &= Diag(2^2, 1^2, 1^2, (.1)^2, 1^2, 1^2, 1^2, 1^2, 1^2) \ , \end{aligned}$$

a more informative prior but strongly biased away from 0 only for the baseline intercept, α_1, and the intercept adjustment for a positive AIDS diagnosis, α_7. These values correspond to our expectation of mean baseline square root CD4 counts of 10 and 7 (100 and 49 on the original scale) for the AIDS-negative and AIDS-positive groups, respectively. This prior forces the drug group intercept (i.e., the effect at baseline) to be very small, since patients were assigned to drug group at random. Indeed, this randomization implies that $\alpha_4 \equiv 0$. Also, the prior allows the data to determine the degree to which CM trajectories depart from horizontal, both before and after the changepoint. Similarly, though α and the β_i are uncorrelated a *priori*, the priors are sufficiently vague to allow the data to determine the proper amount of correlation a *posteriori.*

Since the prior is fully conjugate with the likelihood, the Gibbs sampler is an easily-implemented MCMC method. The required full conditional distributions for this model are similar to those in Example 5.6, and strongly depend on conjugacy and the theoretical results in Lindley and Smith (1972). They are also presented in Section 2.4 of Lange, Carlin, and Gelfand (1992), who used them in a changepoint analysis of longitudinal CD4 cell counts for a group of HIV-positive men in San Francisco. Alternatively, our model could be described graphically, and subsequently analyzed in BUGS (Gilks, Thomas, and Spiegelhalter, 1992; Spiegelhalter, Thomas, Best, and Gilks, 1995a,b). This graphical modeling and Bayesian sampling computer package frees the user from these computational details. See Appendix Subsection C.3.3 for a brief description of graphical modeling and instructions on obtaining the **BUGS** package.

Results

We fit our model by running 5 parallel Gibbs sampler chains for 500 iterations each. These calculations were performed using Fortran 77 with IMSL subroutines on a single Sparc 2 workstation, and took approximately 30 minutes to execute. A crudely overdispersed starting distribution for $\mathbf{V}^{-1}$ was obtained by drawing 5 samples from a Wishart distribution with the R matrix given above but having $\rho = \dim(\beta_i) = 3$, the smallest value for which this distribution is proper. The g^{th} chain for each of the other parameters was initialized at its prior mean plus (g - 3) prior standard deviations, g = 1, . . ., 5. While individually monitoring each of the 1417 parameters is not feasible, Figure 8.2 shows the resulting Gibbs chains for the first 6 components of α, the 3 components of β_8 (a typical random effect), the model standard deviation σ, and the (1,1) and (1,2) elements of $\mathbf{V}^{-1}$, respectively. Also included are the point estimate and 95^{th} percentile of Gelman and Rubin's (1992) scale reduction factor (labeled "G&R" in the figure), which measures between-chain differences and should be close to 1 if the sampler is close to the target distribution, and the autocorrelation at lag 1 as estimated by the first of the 5 chains (labeled "lag 1 acf" in the figure). The figure suggests convergence for all parameters with the possible exception of the two $\mathbf{V}^{-1}$ components. While we have little interest in posterior inference on $\mathbf{V}^{-1}$, caution is required since a lack of convergence

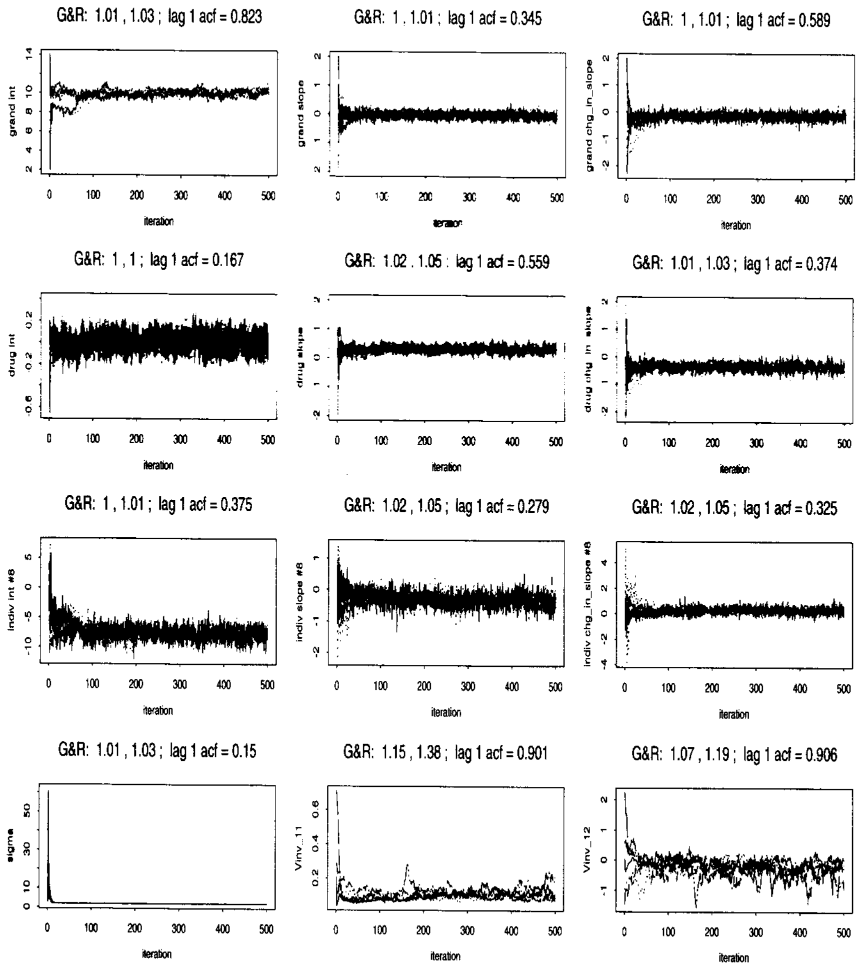


Figure 8.2 *Convergence monitoring plots, five independent chains. Horizontal axis is iteration number; vertical axes are* $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6$, $\beta_{8,1}$, $\beta_{8,2}$, $\beta_{8,3}$, σ, $(\mathbf{V}^{-1})_{11}$, *and* $(\mathbf{V}^{-1})_{12}$.

on its part could lead to false inferences concerning the other parameters that do appear to have converged (Cowles, 1994). In this case, the culprit appears to be the large positive auto correlations present in the chains, a frequent cause of degradation of the Gelman and Rubin statistic in otherwise well-behaved samplers. Since the plotted chains appear reasonably stable, we stop the sampler at this point, concluding that an acceptable degree of convergence has been obtained.

			mode	95% interval	
Baseline	intercept	α_1	9.938	9.319	10.733
	slope	α_2	−0.041	−0.285	0.204
	change in slope	α_3	−0.166	−0.450	0.118
Drug	intercept	α_4	0.004	−0.190	0.198
	slope	α_5	0.309	0.074	0.580
	change in slope	α_6	−0.348	−0.671	−0.074
AIDS Dx	intercept	α_7	−4.295	−5.087	−3.609
	slope	α_8	−0.322	−0.588	−0.056
	change in slope	α_9	0.351	0.056	0.711

Table 8.1 *Point and interval estimates, fixed effect parameters.*

Based on the visual point of overlap of the 5 chains in Figure 8.2, we discarded the first 100 iterations from each chain to complete sampler "burn-in." For posterior summarization, we then used all of the remaining 2000 sampled values despite their autocorrelations, in the spirit of Geyer (1992). For the model variance σ^2, these samples produce an estimated posterior mean and standard deviation of 3.36 and 0.18, respectively. By contrast, the estimated posterior mean of $\mathbf{V}_{11}$ is 13.1, suggesting that the random effects account for much of the variability in the data. The dispersion evident in exploratory plots of the β_i posterior means (not shown) confirms the necessity of their inclusion in the model.

Posterior modes and 95% equal-tail credible intervals (based on posterior mixture density estimates) for all 9 components of the fixed effect parameter vector α are given in Table 8.1. The posterior credible interval for α_4 is virtually unchanged from that specified in its prior, confirming that the randomization of patients to drug group has resulted in no additional information on α_4 in the data. The 95% credible sets of (.074, .580) and (-.671, -.074) for α_5 (pre-changepoint drug slope) and α_6 (post- changepoint drug slope), respectively, suggest that the CD4 trajectories of persons receiving ddI have slopes that are significantly different from those of the ddC patients, both before and after the two-month changepoint. AIDS diagnosis also emerges as a significant predictor of the two slopes and the intercept, as shown in the last three lines in the table.

Plots of the fitted population CD4 trajectories for each of the four possible drug-diagnosis combinations (not shown) suggest that an improvement in square root CD4 count typically occurs only for AIDS-negative patients receiving ddI. However, there is also some indication that the ddC trajectories "catch up" to the corresponding ddI trajectories by the end of the observation period. Moreover, while the trajectories of ddI patients may be

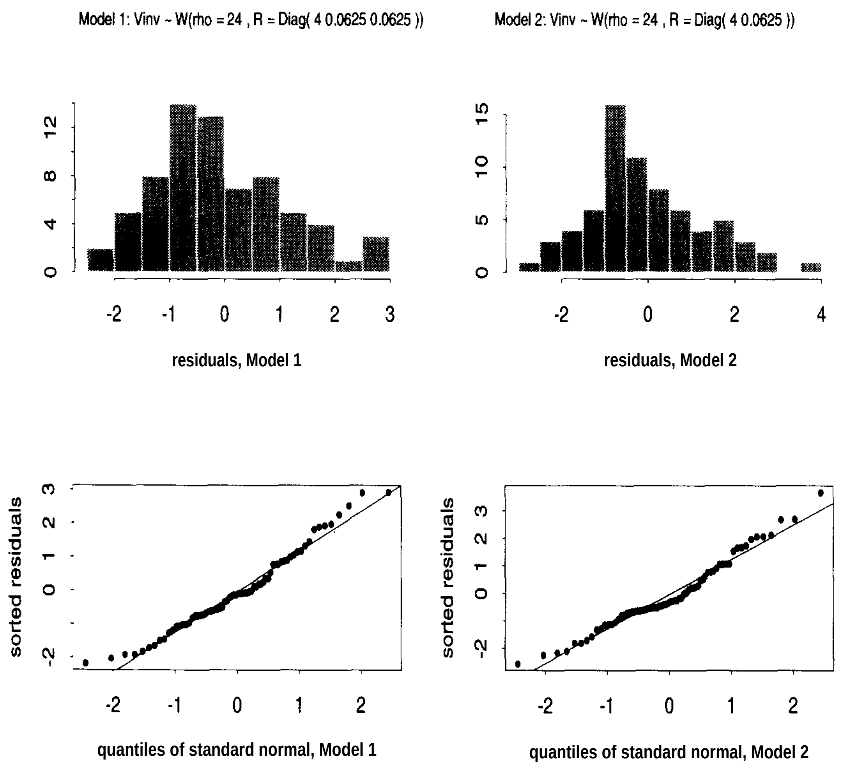


Figure 8.3 *Bayesian model choice based on 70 random observations from the 1405 total. Model 1: changepoint model; model 2: linear decay model.*

somewhat higher, the difference is quite small, and probably insignificant clinically - especially compared to the improvement provided by a negative AIDS diagnosis at baseline.

The relatively small differences between the fitted pre- and post-changepoint slopes suggest that a simple linear decline model might well be sufficient for these data. To check this possibility, we ran 5 sampling chains of 500 iterations each for this simplified model. We compared these two models using the cross-validation ideas described near the beginning of Section 6.5, using a 5% subsample of 70 observations, each from a different randomly selected individual. Figure 8.3 offers a side-by-side comparison of histograms and normal empirical quantile-quantile plots of the resulting residuals for Model 1 (changepoint) and Model 2 (linear). The q-q plots indicate a reasonable degree of normality in the residuals for both models.

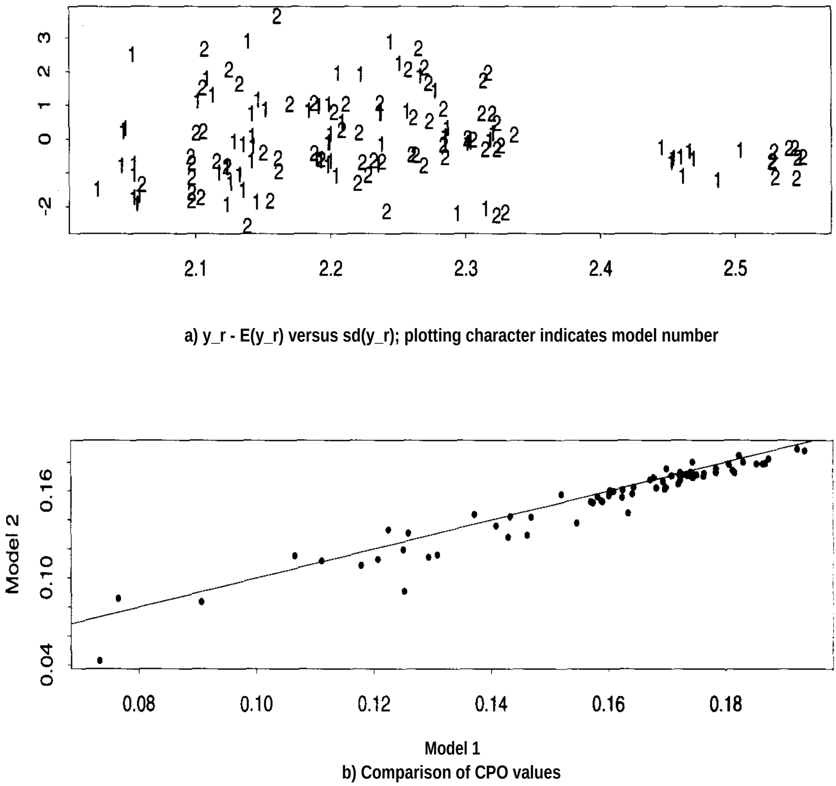


Figure 8.4 *Bayesian model choice based on 70 random observations from the 1405 total. Model 1: changepoint model; model 2: linear decay model.*

The absolute values of these residuals for the two models sum to 66.37 and 70.82, respectively, suggesting almost no degradation in fit using the simpler model, as anticipated. This similarity in quality of fit is confirmed by Figure 8.4(a), which plots the residuals versus their standard deviations. While there is some indication of smaller standard deviations for residuals in the full model (see e.g. the two clumps of points on the right-hand side of the plot), the reduction appears negligible.

A pointwise comparison of CPO is given in Figure 8.4(b), along with a reference line marking where the values are equal for the two models. Since larger CPO values are indicative of better model fit, the predominance of points below the reference line implies a preference for Model 1 (the full model). Still, the log(CPO) values sum to -130.434 and -132.807 for the two models, respectively, in agreement with our previous finding that the improvement offered by Model 1 is very slight. In summary, while the more complicated changepoint model was needed to address the specific clinical questions raised by the trial's designers concerning a CD4 boost at two months, the observed CD4 trajectories can be adequately explained using a simple linear model with our two covariates.

Finally, we provide the results of a more formal Bayesian model choice investigation using the marginal density estimation approach of Subsection 6.3.2 to compute a Bayes factor between the two models. We begin with a trick recommended by Chib and Carlin (1999): for Model 1, we write $p(\{\boldsymbol{\beta}_i\}, \boldsymbol{\alpha}|\sigma^2, \mathbf{V}^{-1}, \mathbf{y}) = p(\{\boldsymbol{\beta}_i\}|\boldsymbol{\alpha}, \sigma^2, \mathbf{V}^{-1}, \mathbf{y}) p(\boldsymbol{\alpha}|\sigma^2, \mathbf{V}^{-1}, \mathbf{y})$, and recognize that the reduced conditional $p(\boldsymbol{\alpha}|\sigma^2, \mathbf{V}^{-1}, \mathbf{y})$ is available in closed form. That is, we marginalize the likelihood over the random effects, obtaining

$$\mathbf{Y}_i|\boldsymbol{\alpha}, \sigma^2, \mathbf{V}, M=1 \sim N\left(\mathbf{X}_i\boldsymbol{\alpha},\ \sigma^2\mathbf{I}_{s_i} + \mathbf{W}_i\mathbf{V}\mathbf{W}_i^\top\right) ,$$

from which we derive the reduced conditional for $\boldsymbol{\alpha}$. This trick somewhat reduces autocorrelation in the chains, and also allows us to work with the appropriate three-block extension of equation (6.14),

$$\begin{aligned}\log f(\mathbf{Y}|M=1) &= \log f(\mathbf{Y}|\boldsymbol{\alpha}', (\sigma^2)', (\mathbf{V}^{-1})', M=1) \\ &\quad + \log p(\boldsymbol{\alpha}', (\sigma^2)', (\mathbf{V}^{-1})'|M=1) - \log p(\boldsymbol{\alpha}'|\mathbf{Y}, (\sigma^2)', (\mathbf{V}^{-1})', M=1) \\ &\quad - \log \hat{p}((\sigma^2)'|\mathbf{Y}, (\mathbf{V}^{-1})', M=1) - \log \hat{p}((\mathbf{V}^{-1})'|\mathbf{Y}, M=1)\end{aligned}$$

for Model 1, and similarly for Model 2.

Using 5 chains of 1000 post-burn-in iterations each, we estimated the log marginal likelihood as -3581.795 for Model 1 and -3577.578 for Model 2. These values are within .03 of values obtained independently for this problem by Chib (personal communication), and produce an estimated Bayes factor of exp[-3577.578- (-3581.795)] = 67.83, moderate to strong evidence in favor of Model 2 (the simpler model without a change point). This is of course consistent with our findings using the less formal model choice tools desribed above.

Summary and discussion

In modeling longitudinal data one must think carefully about the number of random effects to include per individual, and restrict their variability enough to produce reasonably precise estimates of the fixed effects. In our dataset, while there was some evidence that heterogeneous variance parameters σ_i^2 might be needed, including them would increase the number of subject-specific parameters per individual to four - one more than the average number of CD4 observations per person in our study. Fortunately, accidental model overparametrization can often be detected by ultrapoor performance of multichain convergence plots and diagnostics of the variety shown in Figure 8.2. In cases where reducing the total dimension of the model is not possible, recent work by Gelfand, Satin, and Carlin

(1995, 1996) indicates that the relative prior weight assigned to variability at the model's first stage (σ^2) versus its second stage ($\mathbf{V}$) can help suggest an appropriate transformation to improve algorithm convergence. More specifically, had our prior put more credence in small σ^2 values and larger random effects, a sampler operating on the transformed parameters $\alpha^* = \alpha,\ \beta_i^* = \beta_i + \alpha$ would have been recommended.

Missing data are inevitable, and consideration of the possible biasing effects of missingness is very important. Validity of inferences from our models depends on the data being observed at random (i.e., missingness cannot depend on what one would have observed) and an appropriate model structure. Little and Rubin (1987) refer to this combination as "ignorable missingness." In our application, missing CD4 counts are likely the result of patient refusal or a clinical decision that the patient was too ill to have CD4 measured. Thus our data set is very likely biased toward the relatively higher CD4 counts in the later follow-up periods; we are essentially modeling CD4 progression among persons who continue to have CD4 measured. As a check on ignorability, one can explicitly model the missingness mechanism using Tobit-type or logistic models that link missingness to the unobserved data, and compare the resulting inferences. See Cowles et al. (1996) for an application of the Tobit approach in estimating compliance in a clinical trial.

8.1.3 CD4 response to treatment at two months

The longitudinal changepoint model of the previous section takes advantage of all the observations for each individual, but provides only an implicit model for drug-related CD4 response. For an explicit model, we reclassify each patient as "response" or "no response," depending on whether or not the CD4 count increased from its baseline value. That is, for $i = 1, \ldots, m = 367$, the number of patients who had CD4 measured at both the baseline and two month visits, we define $R_i = 1$ if $Y_{i2} - Y_{i1} \geq 0$, and $R_i = 0$ otherwise. Defining p_i as the probability that patient i responded, we fit the probit regression model

$$p_i = \Phi(\gamma_0 + \gamma_1 d_i + \gamma_2 a_i) \; , \tag{8.3}$$

where Φ denotes the cumulative distribution function of a standard normal random variable, the γs are unknown coefficients, and the covariates d_i and a_i describe treatment group and baseline AIDS diagnosis as in Subsection 8.1.2. We used probit regression instead of logistic regression since it is more easily fit using the Gibbs sampler (Carlin and Polson, 1992; Albert and Chib, 1993a).

Under a flat prior on $(\gamma_0, \gamma_1, \gamma_2)'$, the posterior distributions for each of the γ coefficients were used to derive point and interval estimates of the corresponding covariate effects. In addition, given particular values of the covariates, the posterior distributions were used with equation (8.3) to estimate the probability of a CD4 response in subgroups of patients by treatment and baseline prognosis. The point and 95% interval estimates for the regression coefficients obtained using the Gibbs sampler are shown in Table 8.2. Of the three intervals, only the one for prior AIDS diagnosis excludes 0, though the one for treatment group nearly does. The signs on the point estimates of these coefficients indicate that patients without a prior AIDS diagnosis and, to a lesser extent, those in the ddI group were more likely to experience a CD4 response, in agreement with our Subsection 8.1.2 results.

	point estimate	95% confidence limits	
		lower	upper
γ_0 (intercept)	.120	−.135	.378
γ_1 (treatment)	.226	−.040	.485
γ_2 (AIDS diagnosis)	−.339	−.610	−.068

Table 8.2 *Point and interval estimates, CD4 response model coefficients*

Converting these results for the regression coefficients to the probability-of-response scale using equation (8.3), we compare results for a typical patient in each drug-diagnosis group. Table 8.3 shows that the posterior median probability of a response is larger by almost .09 for persons taking ddI, regardless of health status. If we compare across treatment groups instead of within them, we see that the patient without a baseline AIDS diagnosis has roughly a .135 larger chance of responding than the sick patient, regardless of drug status. Notice that for even the best response group considered (baseline AIDS-free patients taking ddI), there is a substantial estimated probability of not experiencing a CD4 response, confirming the rather weak evidence in favor of its existence under the full random effects model (7.10).

	ddI	ddC
AIDS diagnosis at baseline	.502	.413
No AIDS diagnosis at baseline	.637	.550

Table 8.3 *Point estimates, probability of response given treatment and prognosis*

8.1.4 Survival analysis

In order to evaluate the clinical consequences of two-month changes in the CD4 lymphocyte count associated with the study drugs, we performed a parametric survival analysis. As in Subsection 8.1.3, the m = 367 patients with baseline and two-month CD4 counts were divided into responders and non-responders. We wish to fit a proportional hazards model, wherein the hazard function h for a person with covariate values $\mathbf{z}$ and survival or censoring time t takes the form $h(t|\mathbf{z}, \boldsymbol{\beta}) = h_0(t)\exp(\mathbf{z}'\boldsymbol{\beta})$. In our case, we define four covariates as follows: z_0 = 1 for all patients; z_1 = 1 for ddI patients with a CD4 response, and 0 otherwise; z_2 = 1 for ddC patients without a CD4 response, and 0 otherwise; and z_3 = 1 for ddC patients with a CD4 response, and 0 otherwise. Writing $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ and following Cox and Oakes (1984, Sec. 6.1), we obtain the loglikelihood

$$\log L(\boldsymbol{\beta}) = \sum_{i \in \mathcal{U}} \log h(t_i|\mathbf{z}_i, \boldsymbol{\beta}) + \sum_{i=1}^{m} \log S(t_i|\mathbf{z}_i, \boldsymbol{\beta}) \ , \tag{8.4}$$

where S denotes the survival function, $\mathcal{U}$ the collection of uncensored failure times, and $\mathbf{z}_i = (z_{0i}, z_{1i}, z_{2i}, z_{3i})'$. Our parametrization uses nonresponding ddI patients as a reference group; β_1, β_2, and β_3 capture the effect of being in one of the other 3 drug-response groups.

We followed Dellaportas and Smith (1993) by beginning with a Weibull model for the baseline hazard, $h_0(t) = \rho t^{\rho-1}$, but replaced their rejection sampling algorithm based on the concavity of the loglikelihood (8.4) with the easier-to-program Metropolis subchain approach (see Subsection 5.4.4). Our initial concern that the Weibull model might not be rich enough evaporated when we discovered extremely high posterior correlation between ρ and β_0. This suggests the baseline hazard may reasonably be thought of as exponential in our very ill population, and hence we fixed ρ =1 in all subsequent calculations.

The resulting estimated marginal posterior distributions for the βs were fairly symmetric, and those for β_1 and β_2 were centered near 0. However, the 95% equal-tail posterior credible set for β_3, (-1.10, 0.02), suggests predominantly negative values. To ease the interpretation of this finding, we transform the posterior $\boldsymbol{\beta}$ samples into corresponding ones from the survival function using the relation $S(t|\mathbf{z}, \boldsymbol{\beta}) = \exp\{-t \ \exp(\mathbf{z}'\boldsymbol{\beta})\}$. We do this for each drug-response group over a grid of 9 equally-spaced t values from 0 to 800 days. Figure 8.5(a) gives a smoothed plot of the medians of the resulting samples, thus providing estimated posterior survival functions for the four groups. As expected, the group of ddC responders stands out, with substantially improved mortality. A more dramatic impression of this difference is conveyed by the estimated posterior distributions of median survival time, $\theta(\mathbf{z}) = (\log 2)\exp(-\mathbf{z}'\boldsymbol{\beta})$, shown in Figure 8.5(b). While a CD4 response translates into improved survival in both drug groups, the improvement is clinically significant only for ddC recipients.

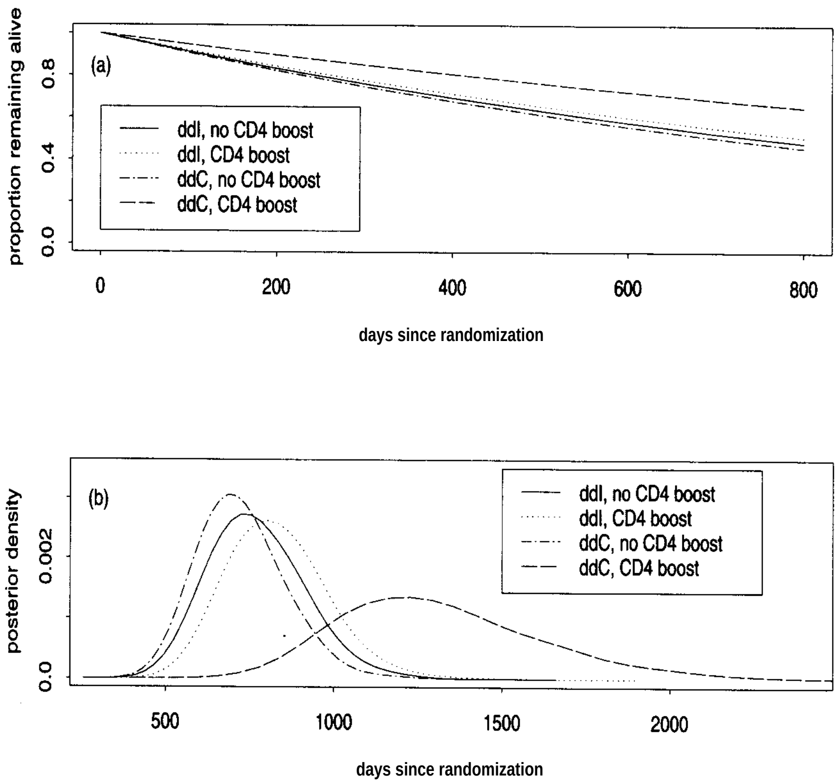


Figure 8.5 *Posterior summaries, ddI/ddC data: (a) estimated survival functions; (b) median survival time distributions.*

8.1.5 Discussion

Analysis of our CD4 data using both a longitudinal changepoint model and a simpler probit regression model suggests that ddC is less successful than ddI in producing an initial CD4 boost in patients with advanced HIV infection. However, this superior short-term performance does not translate into improved survival for ddI patients. In fact, it is the patients "responding" to ddC who have improved survival. It is quite possible that these ddC "responders"' are simply those persons whose CD4 would have increased in the absence of treatment, whereas the ddI "responders" reflect the short-term increase in CD4 produced by ddI. These results provide another example of the caution needed in evaluating treatments on the basis of potential surrogate markers (like CD4 count), and on the use of post-randomization information to explain results. The CD4 count has strong prognostic value, but treatment-induced modifications are not similarly prognostic.

8.2 Robust analysis of clinical trials

8.2.1 Clinical background

In this section we consider another case study related to AIDS, but in the context of the interim monitoring and final analysis of an ongoing clinical trial, namely, the CPCRA toxoplasmic encephalitis (TE) prophylaxis trial. When the degree of immune damage becomes sufficiently severe, an HIV-infected person may develop a specific subset of more than 20 infections, several cancers, a variety of neurological abnormalities including severe declines in mental function, and wasting. Among the most ominous infections is encephalitis due to *Toxoplasma gondii.* This infection is the cause of death in approximately 50% of persons who develop it and the median survival is approximately six months. Additional clinical and immunological background concerning TE is provided in the review paper by Carlin, Chaloner, Louis, and Rhame (1995).

All patients entered into the study had either an AIDS defining illness or a CD4 count of less than 200. In addition all had a positive titre for *toxoplasma gondii* and were therefore at risk for TE. The trial was originally designed with four treatment groups: active clindamycin, placebo clindamycin, active pyrimethamine, and placebo pyrimethamine with an allocation ratio of 2:1:2:1, respectively. It was anticipated that the two placebo groups would be pooled in a final analysis for a three treatment group comparison of placebo, active pyrimethamine, and active clindamycin.

The first patient was randomized in September 1990. The active and placebo clindamycin treatments were terminated soon after in March 1991, due to non-life-threatening toxicities that resulted in the discontinuation of medication for many patients in the active clindamycin group. People on either active or placebo clindamycin were offered rerandomization to either active or placebo pyrimethamine with a 2:1 allocation ratio. The trial then became a simple two treatment study of active pyrimethamine against placebo pyrimethamine.

8.2.2 Interim monitoring

In order to ensure scientific integrity and ethical conduct, decisions concerning whether or not to continue a clinical trial based on the accumulated data are made by an independent group of statisticians, clinicians and ethicists who form the trial's *data safety and monitoring board,* or DSMB. These boards have been a standard part of clinical trials practice (and NIH policy) since the early 1970s; see Fleming (1992) for a full discussion. As a result, the trial's statisticians require efficient algorithms for computing posterior summaries of quantities of interest to the DSMB.

For simplicity, we consider the case of p = 2 covariates: z_{1i} indicates whether individual i received pyrimethamine z_{1i} = 1) or placebo (z_{1i} = 0), and z_{2i} denotes the baseline CD4 cell count of individual i at study entry. Using the method and software of Chaloner et al. (1993), a prior for p_1 was elicted from five AIDS experts, coded as A, B, C, D, and E. Experts A, B, and C are physicians: A practices at a university AIDS clinic, B specializes in the neurological manifestations of AIDS and HIV infection, and C practices at a community clinic. Experts D and E are non-physicians involved in AIDS clinical trials: D manages studies and E is an infectious disease epidemiologist. Each expert's prior was then discretized onto a 31-point grid, each support point having probability 1/31. Finally, each prior was transformed to the β_1 scale and converted to a smooth curve suitable for graphical display using a histogram approach. In what follows we consider only the prior of expert A; the paper by Carlin et al. (1993) provides a more complete analysis. While the trial protocol specified that only the onset of TE was to be treated as an endpoint, expert A was unable to separate the outcomes of death and TE. As such, this prior relates to the dual endpoint of TE or death.

As described in the report by Jacobson et al. (1994), the trial's data safety and monitoring board met on three occasions after the start of the trial in September of 1990 to assess its progress and determine whether it should continue or not. These three meetings analyzed the data available as of the file closing dates 1/15/91, 7/31/91, and 12/31/91, respectively. At its final meeting, the board recommended stopping the trial based on an informal stochastic curtailment rule: the pyrimethamine group had not shown significantly fewer TE events up to that time, and, due to the low TE rate, a significant difference was judged unlikely to emerge in the future. An increase in the number of deaths in the pyrimethamine group was also noted, but this was not a stated reason for the discontinuation of the trial (although subsequent follow-up confirmed this mortality increase). The recommendation to terminate the study was conditional on the agreement of the protocol chairperson after unblinding and review of the data. As a result, the trial did not actually stop until 3/30/92, when patients were instructed to discontinue their study medication.

We now create displays that form a Bayesian counterpart to standard monitoring boundaries. Using the Cox partial likelihood (7.15) as our likelihood, Figure 8.6 diplays the elicited prior, likelihood, exact posterior, and normal approximation to the posterior of β_1 for the four data monitoring dates listed above. The number of accumulated events is shown beneath each display. The exact posterior calculations were obtained by simple summation methods after discretizing β_2 onto a suitably fine grid. Notice that in the first frame the likelihood is flat and the prior equals the exact pos-

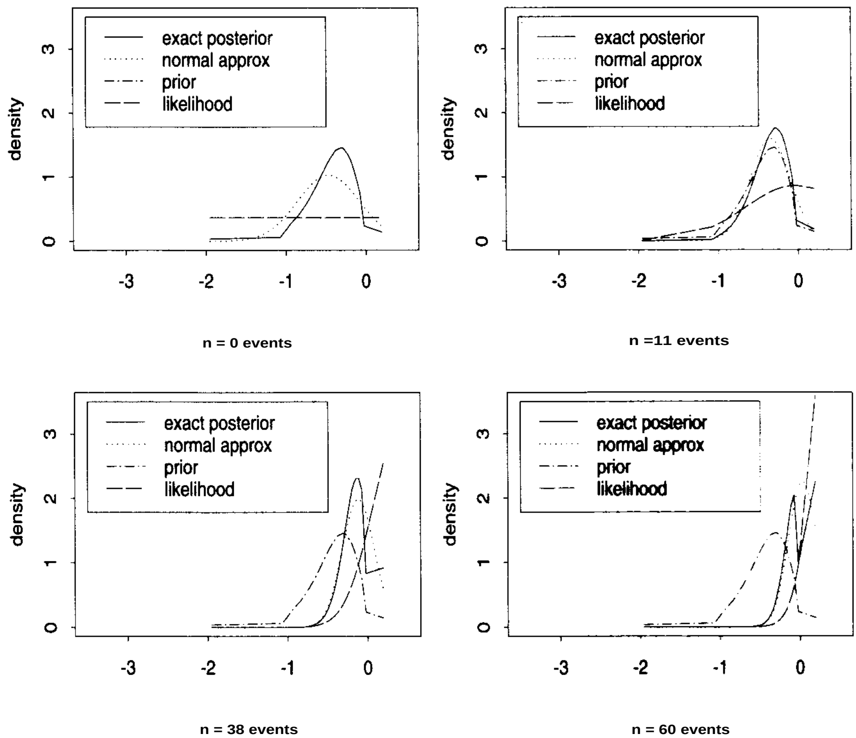


Figure 8.6 *Prior A, likelihood, and posteriors for* β_1 *on four data monitoring dates, TE trial data. Covariate = baseline CD4 count; monitoring dates are 1/15/91, 7/31/91, 12/31/91, and 3/30/92.*

terior, since no information has yet accumulated. From this frame we see that expert A expects the pyrimethamine treatment to be effective at preventing TE/death; only one of the 31 prior support points are to the right of 0. The remaining frames show the excess of events (mostly deaths) accumulating in the treatment group, as the likelihood and posterior summaries move dramatically toward positive β_1 values. The extremely limited prior support for these values (and in fact, no support in the region where the likelihood is maximized) causes the likelihood to appear one-tailed, and the exact posterior to assume a very odd shape. The normal approximation is inadequate when the sample size is small (due to the skew in the prior) and, contrary to the usual theory, also when it is large (due to the truncation effect of the prior on the likelihood).

Next, we turn to the monitoring plot associated with the posteriors in Figure 8.6. The trial protocol specifies a target reduction of 50% in hazard

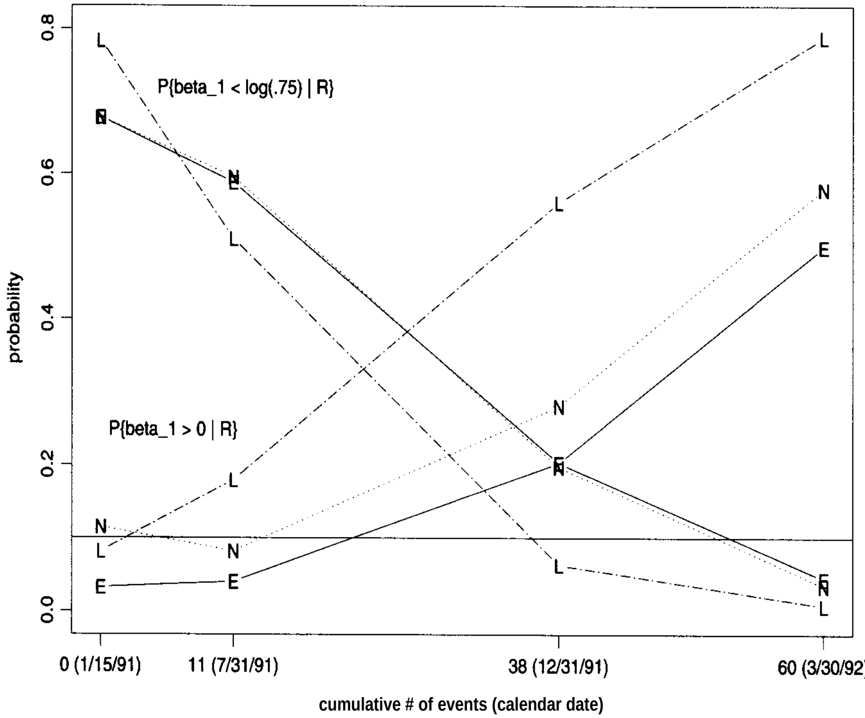


Figure 8.7 *Posterior monitoring plot for* β_1, *TE trial data. E = exact posterior; N = normal approximation; L = likelihood.*

for the treatment relative to control. For a more sensitive analysis we use a 25% reduction as the target, implying the indifference zone boundaries $\beta_{1,U} = 0$ and $\beta_{1,L} = \log(.75) = -.288$. Figure 8.7 plots these two tail areas versus accumulated information for the exact posterior, approximate normal posterior, and likelihood (or more accurately, the posterior under a prior on β_1 that is flat over the support range specified by our expert). This figure nicely captures the steady progression of the data's impact on expert A's prior. Using a stopping level of $p = .10$ (indicated by the horizonta reference line in the figure), the flat prior analysis recommends stopping the trial as of the penultimate monitoring point; by the final monitoring point the data converts the informative prior to this view as well. The inadequacy of the normal approximation is seen in its consistent overstatement of the upper tail probability.

Impact of subjective prior information

Figure 8.7 raises several questions and challenges for further research. For instance, notice that expert A's prior suggests stopping immediately (i.e. not starting the trial) and deciding in favor of the treatment. Though none of the elicitees would have actually recommended such an action (recall the 50% reduction goal), it does highlight the issue of ethical priors for clinical trials, as alluded to earlier. In addition, the fact that expert A's clinically-derived prior bears almost no resemblance to the data, besides being medically interesting, is perhaps statistically troubling as well, since it suggests delaying the termination of an apparently harmful treatment. Perhaps expert A should have received more probabilistic training before elicitation began, or been presented with independent data on the issue.

Prior beliefs common to all five of the experts interviewed included TE being common in this patient population, and a substantial beneficial prophylactic effect from pyrimethamine. However, data from the trial were at odds with these prior beliefs. For instance, out of the 396 people in the trial, only 12 out of 264 on pyrimethamine and only 4 out of 132 on placebo actually got TE. The low rate is believed to be partly due to a change in the standard care of these patients during the trial. During the trial, information was released on the prophylactic efffects of the antibiotic trimethoprim-sulphamethoxazole (TMP/SMX, or Bactrim). TE was extremely rare for patients receiving this drug, and many patients in our study were taking it concurrently with their study medication.

Finally, all five experts put high probability on a large beneficial effect of the treatment, but the rate of TE was actually higher in the pyrimethamine group than in the placebo group. In addition, a total of 46 out of 264 people died in the pyrimethamine group, compared to only 13 out of 132 in the placebo group. Apparently the assumption made at the design stage of the trial that there would be no difference in the non-TE death rates in the different trial arms was incorrect.

Still, the decision makers are protected by the simultaneous display of the flat prior (likelihood) results with those arising from the overly optimistic clinical prior; the results of a skeptical prior might also be included. Overall, this example clearly illustrates the importance of prior robustness testing in Bayesian clinical trial monitoring; this is the subject of Subsection 8.2.3.

Impact of proportional hazards

Our data indicate that the proportional hazards assumption may not hold in this trial. The relative hazard for death in the pyrimethamine group increased over follow-up compared to the placebo group (see Figure 1 of Jacobson et al., 1994). Specifically, while the hazard for the placebo group is essentially constant for the entire study period, the hazard in the pyrimethamine group changes level at approximately six months into the follow-up period. As such, we might reanalyze our data using a model that does not presume proportional hazards. Several authors have proposed models of this type (see e.g. Louis, 1981b; Lin and Wei, 1989; Tsiatis, 1990). Other authors (e.g. Grambsch and Therneau, 1994) create diagnostic plots designed specifically to detect the presence of time-varying covariates in an otherwise proportional hazards setting.

The simplest approach might be to include a parameter in the Cox model (7.15) that explicitly accounts for any hazard rate change. In our case, this might mean including a new covariate z_{p+1} which equals 0 if the individual is in the placebo group $(z_1 = 0)$, equals -1 if the individual is in the drug group $(z_1 = 1)$ and the corresponding event (failure or censoring) occurs prior to time t = 6 months, and equals $+1$ if the individual is in the drug group and the corresponding event occurs after time t = 6. A posterior distribution for β_{p+1} having significant mass far from 0 would then confirm the assumed pattern of nonproportional hazards. Alternatively, we might replace $\beta_1 z_1$ in model (7.15) by $\beta_1 z_1 w(t; \beta_{p+1})$, where w is a weight function chosen to model the relative hazard in the drug group over time. For example, $w(t; \beta_{p+1}) = \exp(\beta_{p+1} t)$ could be used to model an increasing relative hazard over time, with $\beta_1 = 0$ corresponding to the "null" case of proportional hazards.

Simultaneous display of monitoring information based on the flat prior (likelihood) and those based on a variety of priors (the optimistic, the skeptical, and a few personal) protects the monitoring board from over-dependence on a single prior or a collection of mis-calibrated personal priors and indicates the robustness or lack thereof of a monitoring decision. Bayesian monitoring must be robust to be acceptable.

8.2.3 Prior robustness and prior scoping

Phase III clinical trials (like ours) are large scale trials designed to identify patients who are best treated with a drug whose safety and effectiveness has already been reasonably well established. Despite the power and generality of the Bayesian approach, it is unlikely to generate much enthusiasm in monitoring such trials unless it can be demonstrated that the methods possess good frequentist properties. Figure 8.8 displays the posterior for θ under a flat prior (i.e., the standardized marginal likelihood) at each of the four dates mentioned above. Recalling that negative values of θ correspond to an efficacious treatment, the shift toward positive θ values evident in the third and fourth monitoring points reflects an excess of deaths in the treatment group. This emerging superiority of the placebo was quite surprising, both to those who designed the study and to those who provided the prior distributions.

Consider the percentile restriction approach of Subsection 6.1.2. We first integrate β_2 out of the Cox model likelihood $L(\boldsymbol{\beta})$ by a simple summation

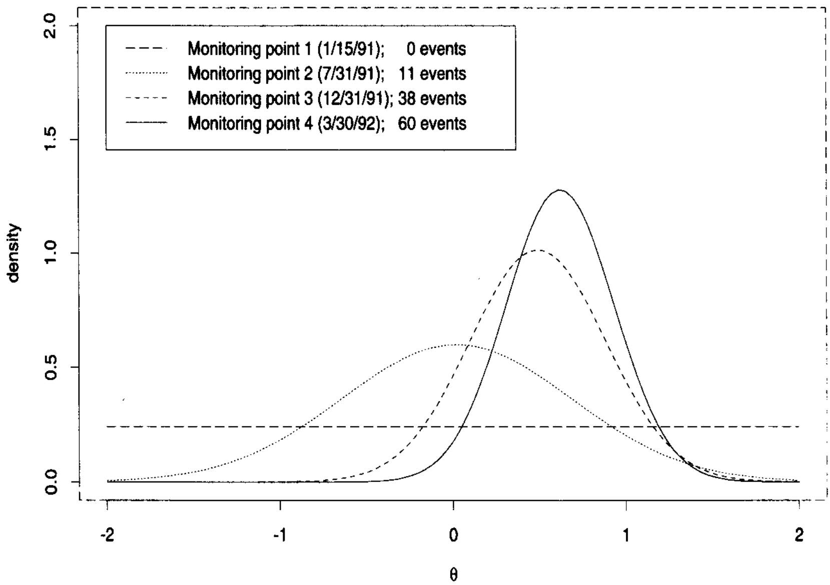


Figure 8.8 *Posterior for the treatment effect under a flat prior, TE trial data. Endpoint is TE or death; covariate is baseline CD4 count.*

method after discretizing β_2 onto a suitably fine grid. This produces the marginal likelihood for β_1, which we write in the notation of Section 6.1.2 simply as $f(x|\theta)$. At the fourth monitoring point, stopping probability p = 0.10 and weight π = 0.25 on the null hypothesis produce the values $f(x|\hat{\theta}_1)$ = 1.28 and $c = 3f(x|0) = 0.48$, where $\hat{\theta}$ = .600 is the marginal MLE of θ.

Suppose we seek the a for which $\mathcal{H}_c$ is nonempty given a specific indifference zone (θ_L, θ_U). Since negative values of θ are indicative of an efficacious treatment, in practice one often takes $\theta_U = 0$ and $\theta_L < 0$, the additional benefit being required of the treatment in order to justify its higher cost in terms of resources, clinical effort, or toxicity. We adopt this strategy and, following our work in the previous subsection, choose $\theta_L = \log(.75) = -.288$, so that a reduction in hazard for the treatment relative to control of at least 25% is required to conclude treatment superiority.

At monitoring point four we have $\theta_U < \hat{\theta} = .600$, so from equation (6.6) the **a** that satisfy the condition $\sup_G \int f(x|\theta)dG(\theta) \geq c$ are such that

$$a_U \geq \frac{[f(x|\theta_U) - f(x|\theta_L)]a_L + [c - f(x|\theta_U)]}{f(x|\hat{\theta}) - f(x|\theta_U)} \, . \tag{8.5}$$

Under the Cox model we obtain $f(x|\theta_L)$ _ .02 and $f(x|\theta_U)$ _ .18, so again taking p = .10 and π = .25, equation (8.5) simplifies to $a_U \geq .145a_L + .273$. For the (a_L, a_U) pairs satisfying this equation, there exists at least one conditional (over H_1 prior G that has these tail areas and permits stopping and rejecting H_0, given the data collected so far. Conversely, for conditional priors G featuring an (a_L, a_U) pair located in the lower region, stopping and rejecting H_0 is not yet possible. Points lying on the boundary between these two regions correspond to the maximum amount of mass the prior may allocate to the indifference zone before stopping becomes impossible.

Next, consider the case of stopping and rejecting H_1 (i.e., accepting H_0) at monitoring point four. This occurs if $P(\theta \neq 0|x) < p$, that is, when $\int f(x|\theta)dG(\theta) \leq c^* = \left(\frac{p}{1-p}\right)\left(\frac{\pi}{1-\pi}\right) f(x|0)$. Hence we are now interested in the infimum of $\int f(x|\theta)dG(\theta)$. Since $\min\{f(x|\theta_L), f(x|\theta_U)\} = f(x|\theta_L) =$ .020 for our dataset, we have from (6.5) that the a satisfying the condition $\inf_G \int f(x|\theta)dG(\theta) \leq c^*$ are such that

$$a_U \geq (1 - 50c^*) - a_L = .700 - a_L \, , \tag{8.6}$$

where we have again let p = .10 and π = .25. Here, increasing π (equivalently c) serves to decrease the intercept of the boundary, and vice versa. In this regard, note that π = 0 is required to make stopping impossible for all **a**. On the other hand, if π is at least .529, then there is at least one prior for which it is possible to stop and accept the null regardless of **a**.

Figure 8.9 shows the four prior tail area regions determined by the two stopping boundaries discussed above. Each region is labeled with an ordered pair, the components of which indicate whether stopping to reject is possible for the prior tail areas in this region under the null and alternative hypotheses, respectively. The relatively large size of the lower left region indicates the mild nature of the evidence provided by these data; there are many prior tail area combinations for which it is not yet possible to stop and reject *either* H_0 or H_1.

Since it may be difficult for the user to think solely in terms of the conditional prior G, we have also labeled the axes of Figure 8.9 with the corresponding unconditional tail areas, which we denote by p_L and p_U. That is, $p_L \equiv P_{unc}$(treatment superior) = $a_L(1-\pi)$, and $p_U \equiv P_{unc}$(control superior) _ $a_U(1-\pi)$; recall that π = .25 in our example. Increasing π (equivalently c) increases only the a_U-intercept in boundary (8.5), so that checking the impact of this assumption is straightforward. In particular, for this dataset we remark that $\pi \leq$ .014 makes stopping and rejecting H_0 possible for all a (boundary (8.5) has a_L-intercept 1), while $\pi \geq$ .410 makes

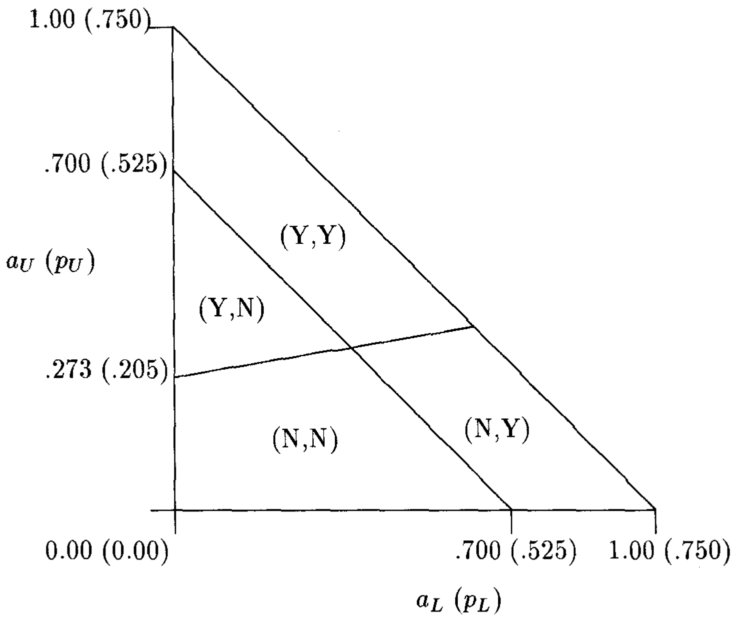
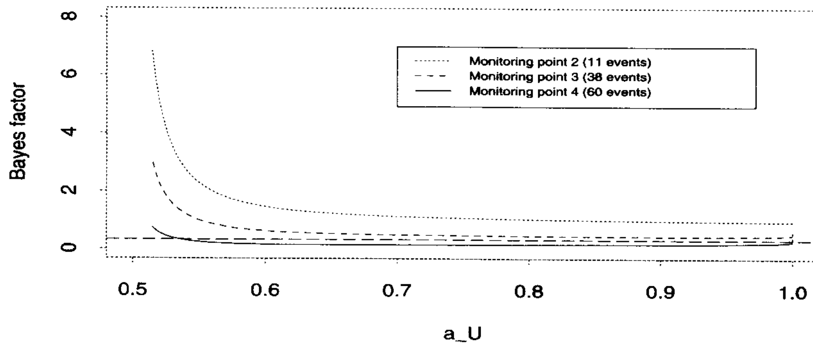


Figure 8.9 *Conditional prior tail area regions, with ordered pair indicating whether it is possible to stop and reject* H_0 *and* $.H_1$, *respectively (Y = yes, N = no).*

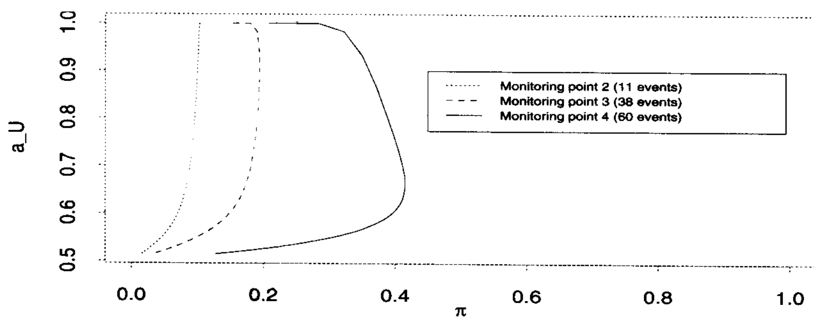
such stopping impossible regardless of a (boundary (8.5) has a_U-intercept 1).

We now turn to the semiparametric approach of Subsection 6.1.2. We use a $N(\theta|\hat{\theta}_k, \hat{\sigma}_k^2)$ approximation to the likelihood at monitoring point k, $k =$ 2, 3, 4, as suggested by Figure 8.8. Since our analysis is targeted toward the members of the Data Safety and Monitoring Board, we restrict attention to "skeptical" conditional priors g by setting $\mu = 0$. We may then compute the Bayes factor BF in favor of H_0 that arises from equations (6.8) and (6.9) as a function of the conditional prior standard deviation τ for $0 < \tau < 4$. To facilitate comparison with previous figures, Figure 8.10(a) plots B not versus τ, but, instead, versus the conditional prior upper tail probability a_U a one-to-one function of τ. Assigning a mass $\pi = .25$ to the null region (the indifference zone) and retaining our rejection threshold of p = .l, we would reject H_0 for B < $\frac{1}{3}$; this value is marked on the figure with a dashed horizontal reference line. The message from this figure is clear: no uniform-normal mixture priors of the form (6.7) allow rejection of H_0 at monitoring points 2 or 3, but by monitoring point 4, rejection of H_0 is favored by all but the most extreme priors (i.e., those that are either extremely vague or essentially point masses at 0).

Next, suppose we wish to investigate robustness as π varies. We reject



a) Bayes factor versus conditional upper tail area



b) Conditional upper tail area versus prior mass on indifference zone;
for combinations to the left of each curve,
all priors result in stopping to reject Ho

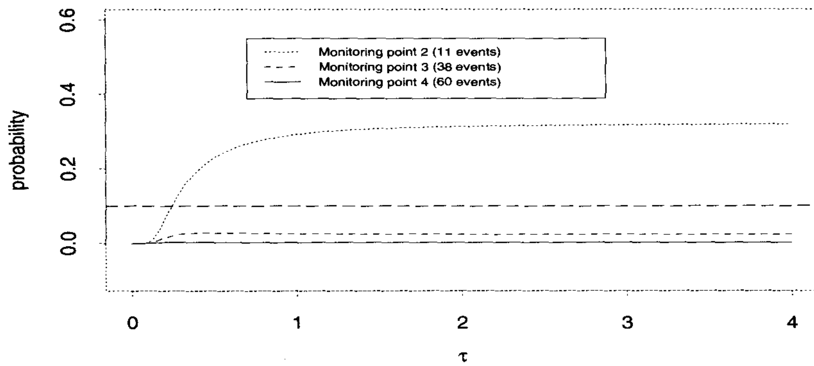
Figure 8.10 *Semiparametric prior partitions, TE trial data.*

H_0 if and only if $B(\tau) \leq \left(\frac{p}{1-p}\right)\left(\frac{1-\pi}{\pi}\right)$, or equivalently,

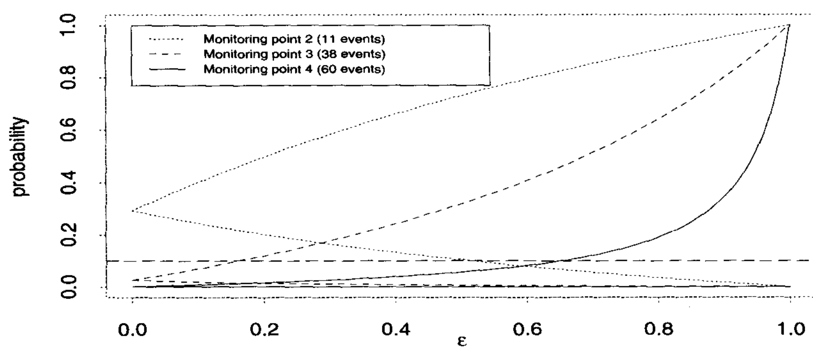
$$\pi \leq \left[\frac{(1-p)B(\tau)}{p} + 1\right]^{-1}$$

Again using the one-to-one relationship between a_U and τ, Figure 8.10(b) plots this boundary in π, a_U -space for the final three monitoring points. Conditional on the data, all priors that correspond to points lying to the left of a given curve lead to rejection of H_0, while all those to the right do not. (Note the difference in interpretation between this graph and Figure 8.9, where each plotted point corresponded to infinitely many priors.) We see the mounting evidence against H_0 in the boundaries' gradual shift to the right, increasing the number of priors that result in rejection.

Finally, we turn to the fully parametric approach outlined in Subsection 6.1.2. Fixing the prior mean μ at 0 again, the prior support for the



a) Posterior probability of H_L versus prior standard deviation



b) Bounds on posterior probability of H_L versus prior contamination factor

Figure 8.11 *Fully parametric prior partitions, TE trial data,*

indifference zone, π, is now determined by τ. This loss of one degree of freedom in the prior means that a bivariate plot like Figure 8.10(b) is no longer possible, but a univariate plot as in Figure 8.10(a) is still sensible. To obtain the strongest possible results for our dataset, in Figure 8.11(a) we plot the posterior probability of $H_L : \theta < \theta_L$, rather than H_0, versus τ. We see that rejection of this hypothesis is possible for all skeptical normal priors by the third monitoring point; by the fourth monitoring point, the evidence against this hypothesis is overwhelming. While this analysis considers a far smaller class of priors than considered in previous figures, it still provides compelling evidence by the third monitoring point that the treatment is not superior in terms of preventing TE or death.

It is of interest to compare our results with those obtained using the e-contamination method of Greenhouse and Wasserman (1995). As mentioned in Subsection 6.1.2, this method enables computation of upper and lower bounds on posterior expectations for priors of the form $(1-\epsilon)g_0+\epsilon g$, where g_0 is some baseline distribution and g can range across the set of all possible priors. Reasonably tight bounds for moderately large ϵ suggest robustness of the original result to changes in the prior. Figure 8.11(b) plots the Greenhouse and Wasserman bounds for the posterior probability of H_L for the final three monitoring points, where we have taken g_0 as our $N(\theta|\mu,\tau^2)$ prior with μ = 0 and τ = 1. At the third monitoring point, the posterior probability is bounded below the rejection threshold of .1 only for ϵ roughly less than .17, suggesting that we cannot stray too far from the baseline normal prior and maintain total confidence in our conclusion. By the fourth monitoring point, however, the upper bound stays below the threshold for ϵ values as large as .65, indicating a substantial degree of robustness to prior specification in our stopping decision.

8.2.4 Sequential decision analysis

In this subsection, we apply the fully Bayesian decision theoretic approach of Subsections 7.7.2 and 7.7.3 to our dataset. The posteriors pictured in Figure 8.8 are nothing but standardized versions of the marginal partial likelihoods for the treatment effect θ given all the data observed up to the given date, and so provide the necessary inputs for these approaches.

Consider first the backward induction approach of Subsection 7.7.2 in this setting. The very near-normal appearance of the curves at the latter three time points justifies an assumption of normality for the prior and likelihood terms. In fact, our dataset provides a good test for the K = 2 calculations in Subsection 7.7.2, since information that would suggest stopping accumulates only at the final two monitoring points. Thus we take the appropriate normal approximation to the 7/31/91 likelihood, a $N(\mu,\sigma^2_{K-2})$ distribution with μ = .021 and $\sigma^2_{K-2} = .664^2$ as our prior $p(\theta)$ and assume independent $N(\theta,\sigma^2_{K-1})$ and $N(\theta,\sigma^2_K)$ distributions for $f(y_{K-1}|\theta)$ and $f(y_K|\theta)$ respectively. Following our earlier discussion we assume the likelihood variance parameters to be known, using the data-based values $\sigma^2_{K-1} = .488^2$ and $\sigma^2_K = .515^2$. Coupled with the normal-normal conjugacy of our prior-likelihood combination, this means that the marginal distributions $m(y_{K-1})$ and $m(y_K|y_{K-1})$ emerge as normal distributions as well. Specifically, the samples required to estimate $L_3^{(K-2)}$ may be obtained as

$$y^*_{K-1,i} \overset{iid}{\sim} N(\mu\,,\ \sigma^2_{K-2}+\sigma^2_{K-1}),\ i=1,\ldots,B,$$

and

$$y^*_{K,ij}|y^*_{K-1,i} \overset{iid}{\sim} N\left(\frac{\sigma^2_{K-1}\mu+\sigma^2_{K-2}y^*_{K-1,i}}{\sigma^2_{K-2}+\sigma^2_{K-1}}\ ,\ \frac{\sigma^2_{K-1}\sigma^2_{K-2}}{\sigma^2_{K-1}+\sigma^2_{K-2}}+\sigma^2_K\right),$$

where j = 1, . . . , B for each i.

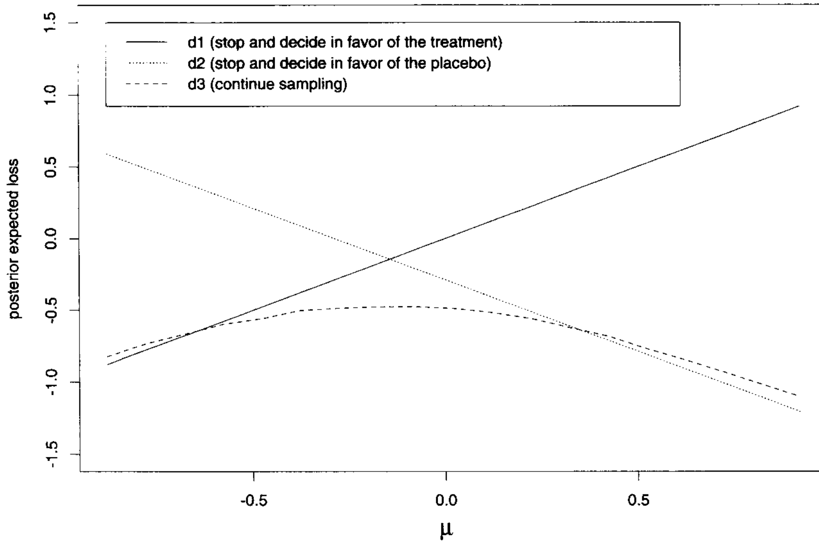


Figure 8.12 *Plot of posterior expected loss $L_i^{(K-2)}$ versus μ for the three decisions (i = 1, 2, 3) computed via backward induction, using the normal/normal model with variances estimated from TE trial data. Choosing the treatment is optimal for μ < -.65, while choosing the placebo is optimal for μ > .36; otherwise, it is optimal to continue the trial to time K - 1.*

We use the same indifference zone as above; in the notation of Subsection 7.7.2 these are $\theta_U = 0$ and $\theta_L = \log(.75) = -.288$. We adopt a very simple loss and cost structure, namely, $s_1^{(k)} = s_2^{(k)} = 1$ and $A_k = .1$, both for all k. Performing the backward induction calculations using B = 5000 Monte Carlo draws, we obtain $L_1^{(K-2)} = .021$, $L_2^{(K-2)} = -.309$, and $L_3^{(K-2)} = -.489$, implying that for this cost, loss, and distributional structure, it is optimal to continue sampling. However, a one-step-back analysis at time K - 1 using $p(\theta|y_{K-1})$ as the prior finds that $L_2^{(K-1)}$ is now the smallest of the three values by far, meaning that the trial should now be stopped and the treatment discarded. This confirms the board's actual recommendation, albeit for a slightly different reason (excess deaths in the drug group, rather than a lack of excess TE cases in the placebo group).

Figure 8.12 expands our time K-2 analysis by replacing the observed 7/31/91 mean $\mu = .021$ with a grid of points centered around this value. We can obtain the intersection points of $L_3^{(K-2)}(\mu)$ with $L_1^{(K-2)}(\mu)$ and $L_2^{(K-2)}(\mu)$ via interpolation as $\gamma_{K-2,L} = -.64$ and $\gamma_{K-2,U} = .36$, respectively. Notice that these values determine a continuation region that contains the observed 7/31/91 mean μ _ .021, and is centered around the midpoint of the indifference zone, $\theta_U + \theta_L)/2$ = -.144.

Since our loss functions are monotonic and we have normal likelihoods, the forward sampling method of Subsection 7.7.3 is applicable and we may compare its results with those above. Using the same loss and cost structure, setting μ = .021, and using $G = 10^6$ Monte Carlo samples, the algorithm quickly deduces that $\gamma_{K-2,L} < \mu < \gamma_{K-2,U}$, so that continuation of the trial is indeed warranted. The algorithm also finds $\gamma_{K-1,L}$ = -0.260 and $\gamma_{K-1,U}$ = -0.044, implying that the trial should be continued beyond the penultimate decision point (12/31/91) only if the posterior mean at this point lies within this interval. Note from Figure 8.8 that the observed mean, $E(\theta|y_{K-1})$ = .491, lies above this interval, again in agreement with the board's decision to stop at this point and reject the treatment.

As with the backward induction method, obtaining exact values via forward sampling for the initial cutoffs $\gamma_{K-2,L}$ and $\gamma_{K-2,U}$ is more involved, since it is again necessary to resort to a grid search of μ values near the solution. Still, using $G = 10^6$ forward Monte Carlo draws we were able to confirm the values found above (-.64 and .36) via backward induction. While the continued need for grid search is a nuisance, we remark that for forward sampling, the *same* G Monte Carlo draws may be used for each candidate μ value.

Using forward sampling we may extend our analysis in three significant ways. First, we extend to the case k = 3, so that the 1/15/91 likelihood now plays the role of the prior, with the remaining three monitoring points representing accumulating data. Second, we return to the "positive part" loss structure (7.16), simply by adding this modification to our definition of the Monte Carlo losses L_j, as in (7.19). Note that this is still a monotone loss specification (hence suitable for forward sampling in the restricted class that depends only on the posterior mean), but one which leads to substantial complications in the backward induction solution. Finally, the positive part loss allows us to order the losses so that incorrect decisions are penalized more heavily the later they are made. Specifically, we take $s_1^{(k)} = s_2^{(k)}$ =k+3, for k=0,1,2,3.

Since forward sampling cannot be implemented with the improper prior suggested by the 1/15/91 likelihood in Figure 8.8, we instead retain a normal specification, i.e., a $N(\mu, \sigma_0^2)$ with μ = 0 and σ_0^2 = 5. This prior is quite vague relative to our rather precise likelihood terms. We also retain our uniform cost structure, A_k = .1 for all k, and simply set $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ = .5602, the sample mean of the three corresponding values estimated from our data, again giving our procedure something of an empirical Bayes flavor. Again using $G = 10^6$ Monte Carlo samples, the forward algorithm quickly determines $\gamma_{0,L} << \mu << \gamma_{0,U}$ (so that the trial should indeed be carried out), and goes on to find precise values for the remainder of the γ vector.

In particular, we obtain $\gamma_{1,L}$ = -0.530 and $\gamma_{1,U}$ = 0.224, an interval that easily contains the observed 7/31/91 mean of .021, confirming the board's decision that the trial should not be stopped this early.

Finally, we investigate the robustness of our conclusions by replacing our normal likelihood with a much heavier-tailed Student's t with ν = 3 degrees of freedom. Since there is now no longer a single sufficient statistic for θ, the optimal rule need not be solely a function of the posterior mean, but we may still use forward sampling to find the best member in this class of posterior mean-based rules. We retain our previous prior, cost, and loss structure, and set $\sigma_k^2 = .560^2/3$, k = 1, 2, 3, so that $Var(Y_k|\theta) = \nu\sigma_k^2/(\nu-2)$ = .5602, as before. Again using $G = 10^6$ Monte Carlo samples, our forward algorithm now requires univariate numerical integration to find the posterior means to compare to the elements of the $\boldsymbol{\gamma}$ vector, but is otherwise unchanged. We now obtain $\gamma_{1,L}$ = -0.503 and $\gamma_{1,U}$ = 0.205, a slightly narrower interval than before (probably since the t_3 has a more concentrated central 95% region than the normal for a fixed variance), but one which still easily contains the observed 7/31/91 mean of .021.

8.2.5 Discussion

As seen in the progression of displays in Subsection 8.2.3, the practical usefulness of a prior partitioning analysis often depends on the breadth of the class of priors considered. One must select the proper class carefully: large classes may lead to overly broad posterior bounds, while narrow classes may eliminate many plausible prior candidates. The quasiunimodal, semiparametric, and fully parametric classes considered here constitute only a small fraction of those discussed in the Bayesian robustness literature. Other possibilities include density bounded classes, density ratio classes, and total variation classes; see Wasserman and Kadane (1992) for a discussion and associated computational strategies.

For computational convenience, we have frequently assumed normal distributions in the above discussion. Using Monte Carlo integration methods, however, prior partitioning can be implemented for any combination of likelihood and prior. Similarly, one could apply the method in settings where θ is a multivariate parameter, such as a multi-arm trial or a simultaneous study of effectiveness and toxicity. As such, prior partitioning offers an attractive complement to traditional prior elicitation and robustness methods in a wide array of clinical trial settings.

Subsection 8.2.4 illustrates the traditional backward induction approach to our two-sided univariate sequential decision problem. It also illustrates that the forward sampling approach dramatically reduces the computational and analytical overhead, thus offering the possibility of fully Bayesian sequential decisionmaking with numerous interim looks, a previously untenable problem. Armed with this more powerful computational framework, many challenges in the practical application of fully Bayesian methods to clincial trial monitoring may now be tackled. Better guidelines for the proper choice of loss functions $l_i^{(k)}(d_i, \theta)$ and corresponding sampling costs $\{A_k\}$, expressed in either dollars earned or lives saved, are needed. For example, if an incorrect decision favoring the treatment were judged C times more egregious than an equivalent one favoring the placebo (i.e., one where the true value lay equally far from the indifference zone but in the opposite direction), we might choose $s_1^{(k)} = C \cdot s_2^{(k)}$. Extensions to more realistic trial settings (beyond the simple normal likelihood and prior illustrated above) would also be welcome.

Finally, though we stress the importance of Bayesian robustness in the monitoring of Phase III clinical trials, more traditional Bayesian analyses are proving effective in the design and analysis of Phase I studies (dose-ranging, toxicity testing) and Phase II studies (indications of potential clinical benefit). In these settings, the ethical consequences of incorrect decisions are less acute than in Phase III studies, and decisions primarily involve the design of the next phase in drug development (including whether there should be another phase!). See Berry (1991, 1993) for discussions of these issues, and Berry et al. (2000) for a case study.

8.3 Spatio-temporal mapping of lung cancer rates

8.3.1 Introduction

Environmental justice and environmental equity are emergent concepts in the development of environmental health policy. These concepts relate to whether exposures to and adverse outcomes from environmental contaminants are shared equitably among various socio-demographic subgroups. Subpopulations of specific interest include low-income groups, racial and ethnic minorities, and sensitive individuals such as children and expectant mothers.

Sound statistical methodology for assessing environmental justice (or detecting environmental "injustices") is needed. The assessment of environmental justice is a complex problem (Wagener and Williams, 1993) with three primary statistical components:

(1) exposure assessment at any location within a geographic region,

(2) accurate estimation of socio-demographic variables across the same geographic region, and

(3) disease (or other outcome of interest) incidence in small regions within the same region.

This case study primarily concerns modeling associations between components (2) and (3).

Several geostatistical smoothing, or *kriging,* methods have been proposed for the analysis of the spatial pattern of disease. In these methods, disease risk is modeled as a continuous surface over the study area. Classical approaches model dependence among observations as part of the single-stage likelihood, while Bayesian approaches typically assume observations to be conditionally independent given the model parameters, and subsequently incorporate dependence at the second stage as part of the prior distribution.

Other approaches estimate rates and counts for administrative districts, usually producing a rate map. A disease map results by color-coding (or greyscale-coding) the estimated rates. Collections of maps for various diseases and/or geographic regions are often published together in a single atlas; see Walter and Birnie (1991) for a comparison of forty-nine disease atlases published during the last twenty years.

Clayton and Kaldor (1987) and Manton et al. (1989) outline empirical Bayes approaches that account for spatial similarities among neighboring and nearby rates. Ghosh (1992) and Devine and Louis (1994) consider constrained empirical Bayes approaches which guard against overshrinkage of the raw rates toward their grand mean. Cressie and Chan (1989) present an exploratory spatial analysis of regional counts of sudden infant death syndrome (SIDS). The regional aspect of the data drives their analysis, which is based on the autoregressive methods for data from a spatial lattice proposed by Besag (1974). Clayton and Bernardinelli (1992) review Bayesian methods for modeling regional disease rates. Besag, York, and Mollié (1991) begin with the Clayton and Kaldor (1987) approach and extend it to a model for separating spatial effects from overall heterogeneity in the rates. Bernardinelli and Montomoli (1992) compare empirical and hierarchical Bayes methods, the latter implemented via MCMC methods. Breslow and Clayton (1993) place the disease mapping problem within the larger framework of generalized linear mixed models and provide approximation schemes for inference.

Here we extend the spatial models described by Besag et al. (1991) to accommodate general temporal effects, as well as space-time interactions. A hierarchical framework for modeling regional disease rates over space and time was given in Subsection 7.8.2. Our model also extends that of Bernardinelli, Clayton et al. (1995), who proposed a particular multiplicative form for space-time interaction. Subsection 8.3.2 presents our dataset of annual county-specific lung cancer rates by sex and race in the state of Ohio for the period 1968-1988, and also outlines the particular form of our model. Computational issues related to these models are presented in Subsection 8.3.3, with special attention paid to parameter identifiability and the associated convergence of our Markov chain Monte Carlo algorithms. The results of fitting our basic model are given in Subsection 8.3.4, along with information on model validation and selection among more parsimonious models. The predictive model choice methods described in Section 6.5 are seen to provide a feasible approach to model comparison in this highly parametrized setting. Finally, we summarize our findings and discuss future implementation and interpretation issues in Subsection 8.3.5.

8.3.2 Data and model description

In this case study, we fit the model described in Subsection 7.8.2 to data originally analyzed by Devine (1992, Chapter 4). Here, y_{ijkt} is the number of lung cancer deaths in county i during year t for gender j and race k in the state of Ohio, and n_{ijkt} is the corresponding exposed population count. These data were originally taken from a public use data tape (Centers for Disease Control, 1988) containing age-specific death counts by underlying cause and population estimates for every county in the United States. Our subset of lung cancer data are recorded for $J = 2$ genders (male and female) and $K = 2$ races (white and nonwhite) for each of the $I = 88$ Ohio counties over an observation period of $T = 21$ years, namely 1968-1988 inclusive, yielding a total of 7392 observations. We obtain internally standardized expected death counts as $E_{ijkt} = n_{ijkt}\bar{y}$, where $\bar{y} = (\sum_{ijkt} y_{ijkt} / \sum_{ijkt} n_{ijkt})$, the average statewide death rate over the entire observation period.

The study area includes the United States Department of Energy Fernald Materials Processing Center (FMPC). The Fernald facility recycles depleted uranium fuel from U.S. Department of Energy and Department of Defense nuclear facilities. The Fernald plant is located in the southwest corner of the state of Ohio approximately 25 miles northwest of the city of Cincinnati. The recycling process creates a large amount of uranium dust. Peak production occurred between 1951 and the early 1960s and some radioactive dust may have been released into the air due to inadequate filtration and ventilation systems. Lung cancer is of interest since inhalation is the primary route of the putative exposure and lung cancer is the most prevalent form of cancer potentially associated with exposure to uranium. We use reported rates for the years 1968-1988 to allow an appropriate (ten- to twenty-year) temporal lag between exposure and disease development.

We apply spatio-temporal models to extend the spatial-only modeling of Devine (1992). The temporal component is of interest to explore changes in rates over a relatively long period of time. Demographic issues (related to environmental justice) are of interest because of possible variation in residential exposures for various population subgroups. In addition, the demographic profile of plant workers and nearby residents most likely evolved over the time period of interest.

Devine (1992) and Devine, Louis, and Halloran (1994) applied Gaussian spatial models employing a distance matrix to the average lung cancer rates for white males over the 21-year period. Xia, Carlin, and Waller (1997) fit Poisson likelihood models incorporating the sex and race covariates to each year separately, thus providing a rough estimate of the change in heterogeneity and clustering present in the data over time, and motivating our spatio-temporal analysis. We begin by fitting a version of model (7.28), namely,

$$\mu_{ijkt} = s_j\alpha + r_k\beta + s_j r_k\xi + \theta_i^{(t)} + \phi_i^{(t)} \,, \tag{8.7}$$

where we adopt the gender and race scores

$$s_j = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases} \quad \text{and} \quad r_k = \begin{cases} 0 & \text{if white} \\ 1 & \text{if nonwhite} \end{cases}$$

Letting $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_I^{(t)})'$, $\boldsymbol{\phi}^{(t)} = (\phi_1^{(t)}, \ldots, \phi_I^{(t)})'$, and denoting the I-dimensional identity matrix by **I**, we adopt the prior structure

$$\boldsymbol{\theta}^{(t)}|\tau_t \overset{ind}{\sim} N\left(\mathbf{0}\,,\, \frac{1}{\tau_t}\mathbf{I}\right) \quad \text{and} \quad \boldsymbol{\phi}^{(t)}|\lambda_t \overset{ind}{\sim} CAR(\lambda_t)\,,\;\; t = 1,\ldots,T\,, \tag{8.8}$$

so that heterogeneity and clustering may vary over time. Note that the socio-demographic covariates (gender and race) do not interact with time or location.

To complete the model specification, we require prior distributions for α, β, ξ, the τ_t and the λ_t. Since $\alpha, \beta,$ and ξ will be identified by the likelihood, we may employ a flat prior on these three parameters. Next, for the priors on the τ_t and λ_t we employed conjugate, conditionally i.i.d. Gamma(a, b) and Gamma(c, d) priors, respectively. As the discussion in Subsection 7.8.2 revealed, some precision is required to facilitate implementation of an MCMC algorithm in this setting. On the other hand, too much precision risks likelihood-prior disagreement. To help settle this matter, we fit a spatial-only (reduced) version of model (8.7) to the data from the middle year in our set (1978, t = 11), using vague priors for λ and τ having both mean and standard deviation equal to 100 ($a = c = 1$, $b = d = 100$). The resulting posterior .025, .50, and .975 quantiles for λ and τ were (4.0, 7.4, 13.9) and (46.8, 107.4, 313.8), respectively. As such, in fitting our full spatio-temporal model (8.7), we retain a = 1, $b = 100$ for the prior on τ, but reset c = 1, d = 7 (i.e., prior mean and standard deviation equal to 7). While these priors are still quite vague, the fact that we have used a small portion of our data to help determine them does give our approach a slight empirical Bayes flavor. Still, our specification is consistent with the advice of Bernardinelli, Clayton, and Montomoli (1995), who suggest that the heterogeneity parameters have prior standard deviation roughly 7/10 that assigned to the clustering parameters. Recasting this advice in terms of prior precisions and the adjacency structure or our CAR prior for the $\phi_i^{(t)}$ we have $\lambda \approx \tau/(2\bar{m})$, where $\bar{m}$ is the average number of counties adjacent to a randomly selected county (about 5-6 for Ohio).

8.3.3 Computational considerations

Due to high dimensionality and complexity of our spatio-temporal model, some form of MCMC algorithm will be needed to obtain estimates of the posterior and predictive quantities of interest. The Gibbs sampler is not well-suited to this problem, however, because, with the exception of the λ_t and τ_t, the full conditional distributions for the model parameters are not standard families. Therefore, we use instead the Metropolis algorithm to obtain the necessary samples. We begin with a univariate version, each parameter having a univariate normal candidate density with mean at the current value of this parameter and variance σ^2 chosen to provide a Metropolis acceptance ratio between 25 and 50% (as per the recommendation of Gelman et al., 1996).

This collection of univariate Metropolis and Gibbs updating steps converges fairly slowly - no surprise in view of the between-parameter posterior correlations we would expect in our model. As we have seen, updating parameters in multivariate blocks (Liu, Wong, and Kong, 1994) is one way to account for these correlations; another is through elementary transformations (Gelfand et al., 1995, 1996). The latter approach is particularly useful here, and is motivated by the fact that the likelihood by itself cannot individually identify the spatial heterogeneity and clustering parameters, but only their sum. For simplicity, we illustrate in the case of a single period of observation ignoring subgroups. We make a linear transformation of variables from $(\boldsymbol{\theta}, \boldsymbol{\phi})$ to $(\boldsymbol{\theta}, \boldsymbol{\eta})$ where $\eta_i = \theta_i + \phi_i, i = 1, \ldots, I$. Writing the posterior on the old scale as $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) \propto L(\boldsymbol{\theta} + \boldsymbol{\phi}; \mathbf{y})p(\boldsymbol{\theta})p(\boldsymbol{\phi})$ the posterior on the new scale is thus $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y}) \propto L(\boldsymbol{\eta}; \mathbf{y})p(\boldsymbol{\theta})p(\boldsymbol{\eta} - \boldsymbol{\theta})$. Along with our conditional independence assumptions, this implies the full conditional distributions

$$p(\eta_i|\eta_{j\neq i}, \boldsymbol{\theta}, \mathbf{y}) \propto L(\eta_i; y_i)\, p(\eta_i - \theta_i|\{\eta_j - \theta_j\}_{j\neq i}) \qquad (8.9)$$

and

$$p(\theta_i|\theta_{j\neq i}, \boldsymbol{\eta}, \mathbf{y}) \propto p(\theta_i)\, p(\eta_i - \theta_i|\{\eta_j - \theta_j\}_{j\neq i}) \qquad (8.10)$$

for $i = 1, \ldots, I$. Since the likelihood informs about η_i directly in (8.9), overall sampler convergence is improved. We also gain the serendipitous side benefit of a closed form (normal) full conditional in (8.10), since the algebraically awkward Poisson likelihood component is no longer present.

8.3.4 Model fitting, validation, and comparison

Using the cycle of univariate Metropolis and Gibbs steps described in Subsection 8.3.3, we ran 5 parallel, initially overdispersed MCMC chains for 500 iterations, a task which took roughly 20 minutes on a Sparc 10 workstation. Graphical monitoring of the chains for a representative subset of the parameters, along with sample autocorrelations and Gelman and Rubin (1992) diagnostics, indicated an acceptable degree of convergence by around the 100^{th} iteration. Using the final 400 iterations from all 5 chains, we obtained the 95% posterior credible sets (-1.10, -1.06), (0.00, 0.05), and (-0.27, -0.17) for α, β, and ξ, respectively. The corresponding point estimates are translated into the fitted relative risks for the four subgroups in Table 8.4. It is interesting that the fitted sex-race interaction ξ reverses the slight advantage white men hold over nonwhite men, making nonwhite females the healthiest subgroup, with a relative risk nearly four times smaller than either of the male groups. Many Ohio counties have very small nonwhite populations, so this result could be an artifact of our inability to model covariate-region interactions.

demographic subgroup	contribution to ε_{jk}	fitted log-relative risk	fitted relative risk
white males	0	0	1
white females	α	−1.08	0.34
nonwhite males	β	0.02	1.02
nonwhite females	$\alpha + \beta + \xi$	−1.28	0.28

Table 8.4 *Fitted relative risks for the four socio-demographic subgroups in the Ohio lung cancer data.*

Turning to the spatio-temporal parameters, histograms of the sampled values (not shown) showed $\theta_i^{(t)}$ distributions centered near 0 in most cases, but $\phi_i^{(t)}$ distributions typically removed from 0. This suggests some degree of clustering in the data, but no significant additional heterogeneity beyond that explained by the CAR prior. Figures 8.13 and 8.14 investigate this issue a bit further by checking for differential heterogeneity and clustering effects over time. For example, Figure 8.13 plots the estimated posterior medians for the clustering parameters λ_t versus t. The solid line in the plot is the least squares regression fit, while the dashed line is the result of a Tukey-type running median smoother (smooth in *S-plus*). A clear, almost linear increase is observed, suggesting that the spatial similarity of lung cancer cases is increasing over the 21-year time period. On the other hand, the posterior medians for τ_t plotted versus t in Figure 8.14 are all quite near the prior mean of 100 (again suggesting very little excess heterogeneity) and provide less of an indication of trend. What trend there is appears to be downward, suggesting that heterogeneity is also increasing over time (recall that τ_t is the precision in a mean-zero prior for $\theta_i^{(t)}$).

Since under our model the expected number of deaths for a given subgroup in county i during year t is $E_{ijkt}\exp(\mu_{ijkt})$, we have that the (inter-

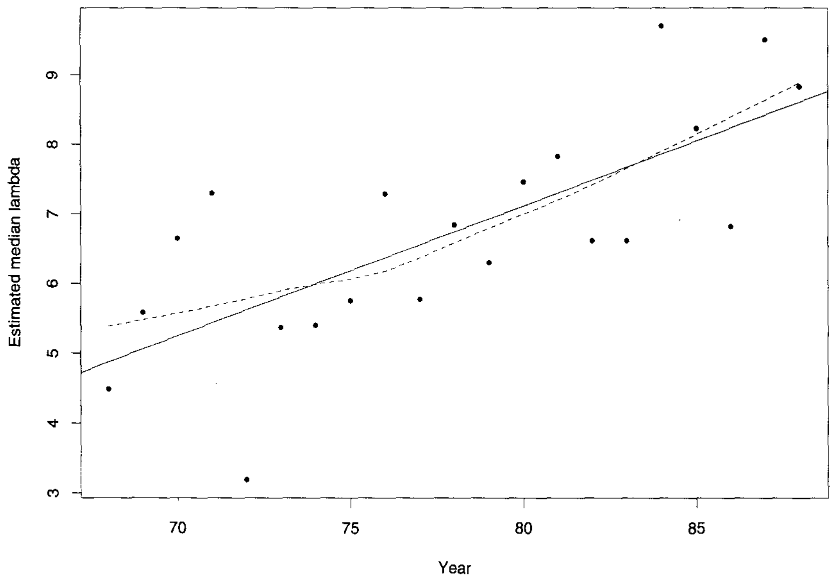


Figure 8.13 *Estimated posterior medians for λ_t versus t, full model.*

nally standardized) expected death rate per thousand is $1000\bar{y}\exp(\mu_{ijkt})$. The first row of Figure 8.15 maps point estimates of these fitted rates for nonwhite females during the first (1968), middle (1978), and last (1988) years in our dataset. These estimates are obtained by plugging in the estimated posterior medians for the μ_{ijkt} parameters calculated from the output of the Gibbs sampler. The rates are greyscale-coded from lowest (white) to highest (black) into seven intervals: less than .08, .08 to .13, .13 to .18, .18 to .23, .23 to .28, .28 to .33, and greater than .33. The second row of the figure shows estimates of the variability in these rates (as measured by the interquartile range) for the same subgroup during these three years. These rates are also greyscale-coded into seven intervals: less than .01, .01 to .02, .02 to .03, .03 to .04, .04 to .05, .05 to .06, and greater than .06.

Figure 8.15 reveals several interesting trends. Lung cancer death rates are increasing over time, as indicated by the gradual darkening of the counties in the figure's first row. But their variability is also increasing somewhat, as we would expect given our Poisson likelihood. This variability is smallest for high-population counties, such as those containing the cities of Cleveland (northern border, third from the right), Toledo (northern border, third from the left), and Cincinnati (southwestern corner). Lung cancer rates are high in these industrialized areas, but there is also a pattern of generally increasing rates as we move from west to east across the state for a given year. One possible explanation for this is a lower level of smoking among

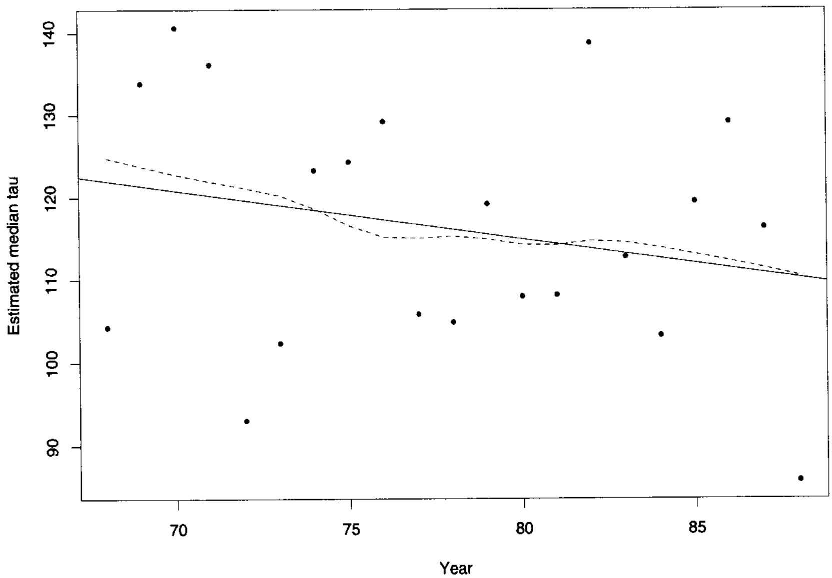


Figure 8.14 *Estimated posterior medians for* τ_t *versus t, full model.*

persons living in the predominantly agricultural west, as compared to those in the more mining and manufacturing-oriented east. Finally, we note that while our conclusions of increasing clustering *and* increasing heterogeneity over time may have at first glance seemed contradictory, they are confirmed by the fitted death rate maps. We see increasing evidence of clustering among the high rate counties, but with the higher rates increasing and the lower rates remaining low (i.e., increasing heterogeneity statewide). Again, since the higher rates tend to emerge in the poorer, more mountainous eastern counties, we have an indication of decreasing environmental equity over the last few decades.

Regarding the possible impact of the Fernald facility, Figure 8.16 shows the fitted lung cancer death rates per 1000 population by year for white males in Hamilton county (which contains the facility), as well as Butler, Warren, and Clermont counties (which are adjacent to Hamilton). The statewide white male death rate is also plotted by year for comparison. (We focus on white male mortality since this is the demographic group most likely to have been affected at the facility during the potential exposure window in the 1950s.) We observe substantially elevated rates of lung cancer in Hamilton county. Rates in the adjacent counties are similar in magnitude to the statewide rate, but do seem to be increasing a bit more rapidly. In addition, the adjacent county with the highest rates overall is Butler, which is immediately north of Hamilton and second-closest to Fernald.

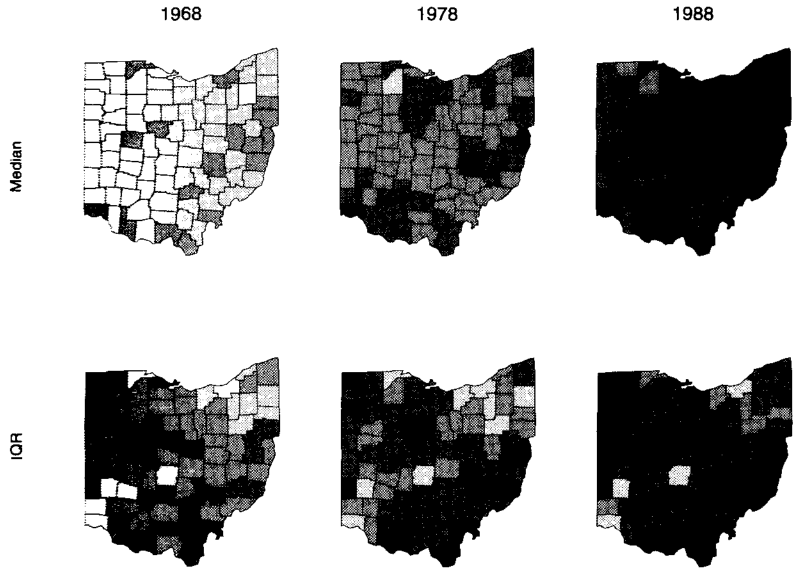


Figure 8.15 *Posterior median and interquartile range (IQR) by county and year, nonwhite female lung cancer death rate per 1000 population (see text for greyscale key).*

Model Validation, comparison, and selection

The results of the previous subsection suggest that our full model (8.7) is overfitting the data, especially with respect to heterogeneity over time. As such. we now contemplate several reduced models, judging their fit relative to the full model using the expected predictive deviance (EPD) score (6.32). These models and their scores are shown in Table 8.5, along with the partitioning of the EPD into its likelihood ratio statistic (LRS) and predictive variability penalty (PEN) components. The first row of the table shows the full model and its scores; the second and third rows show scores for the two models that eliminate the entire collection of spatial and heterogeneity effects, respectively. Not surprisingly, the latter change leads to no significant increase in the overall score, but the former change also seems to have only a small impact (the EPD scores for these three models are essentially equivalent to the order of accuracy in our MCMC calculations). Apparently, despite being centered around a common mean κ_θ *a priori*, the $\theta_i^{(t)}$ parameters are able to adapt their posterior levels to account for the increasing cancer rate over time.

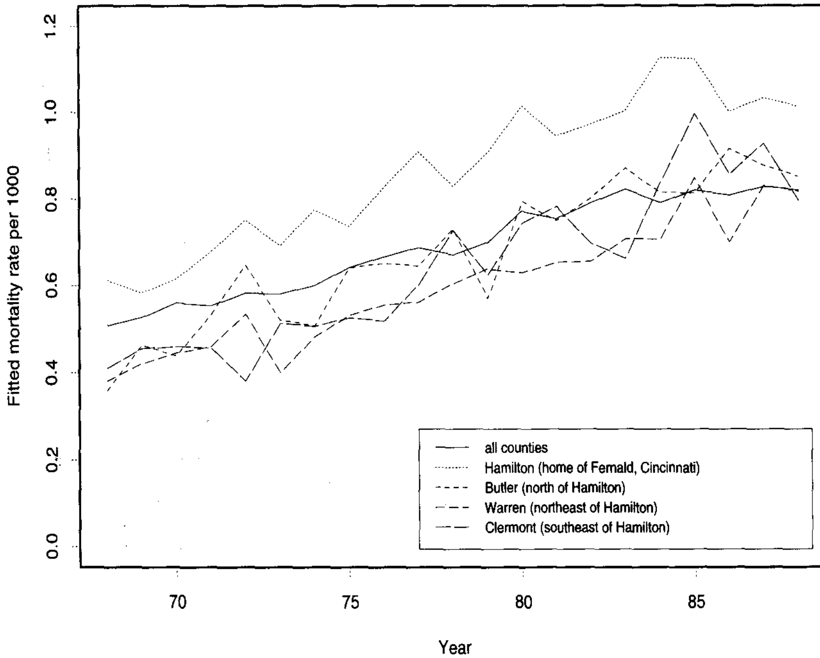


Figure 8.16 *Estimated white male lung cancer death rates per 1000 by year, statewide average and counties near the Fernald Materials Processing Center.*

The next row of Table 8.5 considers the model

$$\mu_{ijkt} = \varepsilon_{jk} + \delta_t + \phi_i \; , \tag{8.11}$$

which eliminates the excess heterogeneity and space-time interaction effects, but still allows a general form for the overall time trend, subject to

model for μ_{ijkt}	LRS	PEN	EPD
$\varepsilon_{jk} + \theta_i^{(t)} + \phi_i^{(t)}$	5374.87	5806.07	11180.94
$\varepsilon_{jk} + \theta_i^{(t)}$	4941.32	6017.06	10958.39
$\varepsilon_{jk} + \phi_i^{(t)}$	5180.25	5808.73	10988.98
$\varepsilon_{jk} + \delta_t + \phi_i$	7274.70	5725.19	12999.89
$\varepsilon_{jk} + \gamma t$	9444.20	4756.31	14200.50

Table 8.5 *Likelihood ratio statistics (LRS), predictive variability penalties (PEN), and overall expected predicted deviance (EPD) scores, spatio-temporal Ohio lung cancer models.*

an identifiability constraint such as $\sum_t \delta_t = 0$. We implemented this constraint numerically at the end of each iteration by recentering the sampled δ_t values around 0 and also adjusting the ϕ_i values accordingly. The prior for the of is a similar, simplified version of that in (8.8) above, namely $\phi_i \sim CAR(\lambda)$. Despite the smaller penalty associated with this more parsimonious model, the increase in the LRS leads to a larger (i.e. poorer) overall EPD score. This increase in score over the first three models is "significant" in the sense that it reoccurred in five other, independent MCMC runs.

Finally. the last row of the table eliminates the spatial aspect of the model altogether. and specifies a linear form for the temporal trend. The resulting model has no random effects and only 5 parameters in total. Again we see that the decrease in the penalty term fails to offset the reduction in quality of fit. resulting in a further significant increase in the overall EPD score. Of course. many other spatio-temporal models might well be considered, such as other functional forms or perhaps an AR(1) structure for the δ_t, multiplicative forms for the spatio-temporal interaction, and so on.

8.3.5 Discussion

With regard to the impact of the Fernald plant, our results are consistent with those obtained by Devine (1992) using a Gaussian spatial-only model. However. they are difficult to interpret due to two confounding factors. First. Hamilton county is also home to Cincinnati, a large urban area. We have already seen in Figure 8.15 that elevated rates are often present in such areas. Second, our model does not incorporate information on smoking prevalence, the most important risk factor for lung cancer. While we currently have no county- or year-specific information on these two factors, they could conceivably be accounted for using census information (say, using population density and the number of cigarettes sold per capita, respectively).

The difficulty in assessing the impact of Fernald points to two larger issues in the investigation of environmental equity. First, it illustrates the problem of determining "clusters" of disease and their "significance." In our dataset. Fernald was known in advance to be a possible source of contamination, so searching for high cancer rates in its vicinity was justified. However. individuals often scan maps like ours looking for "hot spots," and subsequently launch investigations to discover the "cause" of the unusually high rates. Of course, "data-snooping" is bound to uncover such hot spots, many of which will have no practical significance. Still, the detection of disease clusters in the absence of prior information is a legitimate goal, and one which has a long history in the statistical literature (see e.g. Diggle, 1983, and Waller et al., 1994). Here again, the full propagation of uncertainty throughout all stages of the analysis (and associated smoothing of outlying

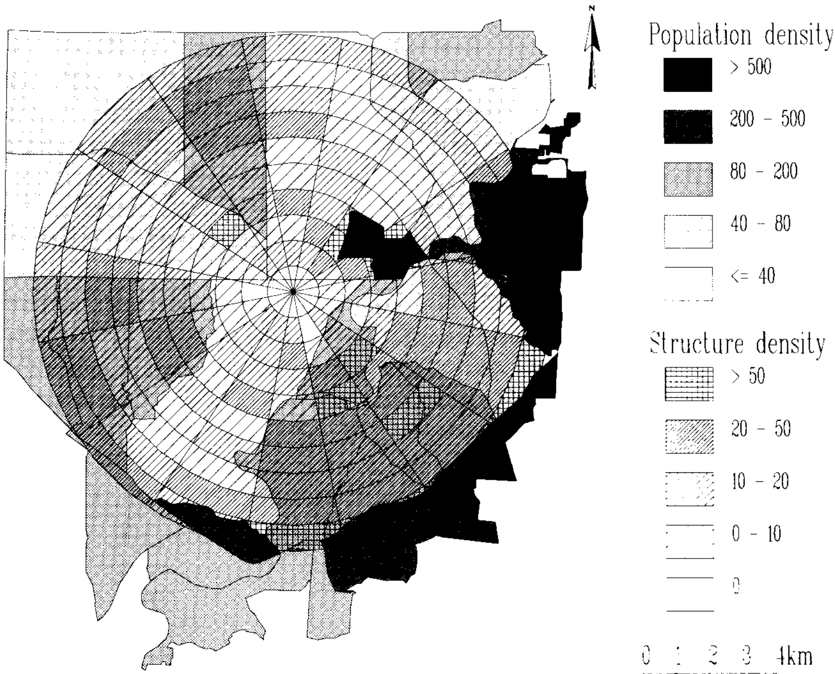


Figure 8.17 *Census block groups and 10-km windrose near the FMPC site, with 1990 population density by block group and 1980 structure density by cell (both in counts per* km^2 *).*

rates) afforded by the Bayesian approach suggests it as most appropriate for attacking this difficult problem.

A second general issue raised by Fernald is that of *misalignment* between the three data sources involved: exposure, disease incidence, and socio-demographic covariates. Exposures are typically available either by geographic regions within a particular assessment domain, or perhaps at certain precise geographic coordinates - say, corresponding to monitoring stations or a reported toxic release. Disease incidence is instead typically available only as summary counts or rates over geographic regions (like our county-wide Ohio cancer rates). Socio-demographic covariates are also usually available as summary counts, but they may be defined over different geographic regions (e.g., census tracts or zip codes). Computer-based *geographic information systems* (GIS), software programs that blend color mapping with modern database systems, are extremely effective tools for handling data misalignment, but they serve only as descriptive tools for summarizing the data, and do not at present allow statistical inference.

The GIS display shown in Figure 8.17 illustrates this issue in the context of our Fernald case study, displaying information on two geographically misaligned regional grids. The first is an exposure "windrose," consisting of 10 concentric circular bands at 1-kilometer radial increments divided into 16 compass sectors (N, NNW, NW, WNW, W, etc.). Using a risk model that accounts for air and groundwater flows near the plant, Killough et al. (1996) provide cumulative radon exposure estimates for each of the 160 cells in the windrose by gender and quinquennial age group (0-4, 5-9, . . . , 35-39) and quinquennial time period (1950-1954, ..., 1985-1989). If we knew the number of persons at risk in each of these cell-gender-age-time categories, we could simply multiply them by the corresponding risk estimates and sum the results to produce expected numbers of observed lung cancers due to plant exposure.

Unfortunately, the source of information for persons at risk is what gives rise to the second regional grid shown in Figure 8.17. These irregularly shaped regions are the boundaries of 39 U.S. Census Bureau block groups, for which 1990 population counts are known. Population density (persons per km^2 is shown using greyscale shading in the figure. The intersection of the two non-nested zonation systems results in 389 regions called atoms, which can be aggregated appropriately to form either cells or block groups.

In order to reallocate the population counts from the block group grid to the exposure cell grid, we may take advantage of a covariate available on the latter. This is the cell-specific residential structures count, obtained by overlaying the windrose onto U.S. Geological Survey (USGS) aerial maps. Structure density (also in counts per km^2) is shown using hatching in Figure 8.17. Using this covariate, Bayesian model-based allocation of the population counts from the block groups to the atoms (and subsequently reaggregated to the cells) should perform better than a simple area-based interpolation, as well as allow full posterior inference (rather than merely point estimates).

Figure 8.18 shows the resulting estimated population density (posterior medians) by cell. Without going into detail here, the approach uses conditionally independent Poisson-multinomial models that incorporate the structure covariate as well as regional areas and the known constraints on the atom-level structure counts in a single cell (or the atom-level population counts in a single block group). Spatial correlation is modeled using a CAR prior structure, and age- and gender-specific allocations are also available. Of special concern are the spatial "edge zones," atoms resulting from block groups extending outside the windrose (and thus for which no structure covariate information is available). See Mugglin et al. (2000) for modeling and implementational details. See also Wolpert and Ickstadt (1998) for more on Bayesian methods for handling data misalignment, and Best et al. (2000) for a related case study in childhood respiratory illness.

Figure 8.18 *Imputed population densities (persons/km^2) by cell for the FMPC windrose.*

APPENDIX A

Distributional catalog

This appendix provides a summary of the statistical distributions used in this book. For each distribution, we list its abbreviation, its support, constraints on its parameters, its density function, and its mean, variance, and other moments as appropriate. We also give additional information helpful in implementing a Bayes or EB analysis (e.g., relation to a particular sampling model or conjugate prior). We denote density functions generically by the letter p, and the random variable for the distribution in question by a Roman letter: x for scalars, x for vectors, and V for matrices. Parameter values are assigned Greek letters. The reader should keep in mind that many of these distributions are most commonly used as prior distributions, so that the random variable would itself be denoted as a Greek character.

In modern applied Bayesian work, a key issue is the generation of random draws from the distributions in this catalog. Since whole textbooks (Devroye, 1986; Ripley, 1987; Gentle, 1998) have been devoted to this subject, we make no attempt at a comprehensive treatment here, but only provide a few remarks and pointers. First, many modern Bayesian software packages (such as BUGS and others described in Appendix Section C.3) essentially "hide" the requisite sampling from the user altogether. High-level (but still general purpose) programming environments (such as S-Plus, Gauss, or MOW) feature functions for generating samples from most of the standard distributions below. Lower-level languages (such as Fortran or C) often include only a $Uniform(0, 1)$ generator, since samples from all other distributions can be built from these. In what follows, we provide a limited number of generation hints, and refer the reader to the aforementioned textbooks for full detail.

A.1 Discrete

A.1.1 Univariate

Binomial: $X \sim Bin(n, \theta)$, $x = 0, 1, 2, \ldots, n$, $0 \leq \theta \leq 1$, n any positive integer, and

$$p(x|n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} ,$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} .$$

$E(X) = n\theta$ and $Var(X) = n\theta(1-\theta)$.

The binomial is commonly assumed in problems involving a series of independent, success/failure trials all having probability of success θ *(Bernoulli trials):* X *is* the number of successes in *n* such trials. The binomial is also the appropriate model for the number of "good" units obtained in *n* independent draws with replacement from a finite population of size *N* having *G* good units, and where θ = *G/N*, the proportion of good units.

Poisson: $X \sim Po(\theta)$, $x = 0, 1, 2, \ldots$, $\theta > 0$, and

$$p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$$

$E(X) = \theta$ and $Var(X) = \theta$.

The Poisson is commonly assumed in problems involving counts of "rare events" occurring in a given time period, and for other discrete data settings having countable support. It is also the limiting distribution for the binomial when n goes to infinity and the expected number of events converges to θ.

- Negative Binomial: $X \sim NegBin(r, \theta)$, $x = 0, 1, 2, \ldots$, $0 \leq \theta \leq 1$, r any positive integer, and

$$p(x|r, \theta) = \binom{x+r-1}{x} \theta^r (1-\theta)^x .$$

$E(X) = r(1-\theta)/\theta$ and $Var(X) = r(1-\theta)/\theta^2$.

The name "negative binomial" is somewhat misleading; "inverse binomial" might be a better name. This is because it arises from a series of Bernoulli trials where the number of *successes,* rather than the total number of trials, is fixed at the outset. In this case, X is the number of failures preceding the r^{th} success. This distribution is also sometimes used to model countably infinite random variables having variance substantially larger than the mean (so that the Poisson model would be inappropriate). In fact, it is the marginal distribution of a Poisson random variable whose rate θ follows a gamma distribution (see Section A.2 below).

- Geometric: $X \sim Geom(\theta) \equiv NegBin(1, \theta)$, $x = 0, 1, 2, \ldots$, $0 \leq \theta \leq 1$, and

$$p(x|\theta) = \theta(1-\theta)^x .$$

$E(X) = (1-\theta)/\theta$ and $Var(X) = (1-\theta)/\theta^2$.

The geometric is the special case of the negative binomial having $r = 1$; it is the number of failures preceding the first success in a series of Bernoulli trials.

A.1.2 Multivariate

9 Multinomial: $\mathbf{X} \sim Mult(n, \boldsymbol{\theta})$ for $\mathbf{X} = (x_1, \ldots, x_k)'$, where $x_i \in \{0, 1, 2, \ldots, n\}$ and $\sum_{i=1}^k x_i = n$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ where $0 \leq \theta_i \leq 1$ and $\sum_{i=1}^k \theta_i = 1$, and

$$p(\mathbf{x}|n, \boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k \theta_i^{x_i} .$$

$E(X_i) = n\theta_i$, $Var(X_i) = n\theta_i(1-\theta_i)$, and $Cov(X_i, X_j) = -n\theta_i\theta_j$.

The multinomial is the multivariate generalization of the binomial: note that for $k = 2$, $Mult(n, \boldsymbol{\theta} = Bin(n, \theta$ with $\theta = \theta_1 = 1 - \theta_2$. For a population having $k > 2$ mutually exclusive and exhaustive classes, X_i is the number of elements of class i that would be obtained in n independent draws with replacement from the population. Note well that this is really only (k-1)-dimensional distribution, since $x_k = n - \sum_{i=1}^{k-1} x_i$ and $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$.

A.2 Continuous

A.2.1 Univariate

- Beta: $X \sim Beta(\alpha, \beta)$, $x \in [0, 1]$, $\alpha > 0$, $\beta > 0$, and

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} ,$$

where $\Gamma(\cdot)$ denotes the *gamma function,* defined by the integral equation

$$\Gamma(\alpha) \equiv \int_0^\infty y^{\alpha-1} e^{-y} dy , \quad \alpha > 0 .$$

Using integration by parts, we have that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, so since $\Gamma(1) = 1$, we have that for integer α,

$$\Gamma(\alpha) = (\alpha - 1)! .$$

One can also show that $\Gamma(1/2) = \sqrt{\pi}$.

For brevity, the pdf is sometimes written as

$$\frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} ,$$

where $B(\cdot, \cdot)$ denotes the *beta function,*

$$B(\alpha, \beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} .$$

Returning to the distribution itself, we have $E(X) = \alpha/(\alpha+\beta)$ and $Var(X) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$; it is also true that $E(X^2) = \alpha(\alpha+1)/[(\alpha+\beta)(\alpha+\beta+1)]$. The beta is a common and flexible distribution for continuous random variables defined on the unit interval (so that rescaled and/or recentered versions may apply to any continuous variable having finite range). For $\alpha < 1$, the distribution has an infinite peak at 0; for $\beta < 1$, the distribution has an infinite peak at 1. For $\alpha = \beta = 1$, the distribution is flat (see the next bullet point). For α and β both greater than 1, the distribution is concave down, and increasingly concentrated around its mean as $\alpha + \beta$ increases. The beta is the conjugate prior for the binomial likelihood; for an example, see Subsection 2.3.4.

Uniform: $X \sim Unif(\theta_L, \theta_U)$, $x \in [\theta_L, \theta_U]$, $-\infty < \theta_L < \theta_U < \infty$, and

$$p(x|\theta_L, \theta_U) = \frac{1}{\theta_U - \theta_L} .$$

$E(X) = (\theta_L + \theta_U)/2$ and $Var(X) = (\theta_U - \theta_L)^2/12$. Clearly $Unif(0,1) \equiv Beta(1,1)$.

- Normal (or Gaussian): $X \sim N(\mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$E(X) = \mu$ and $Var(X) = \sigma^2$. The normal is the single most common distribution in statistics, due in large measure to the Central Limit Theorem. It is symmetric about *p,* which is not only its mean, but its median and mode as well.

A $N(0, 1)$ variate Z can be generated as $\sqrt{-2\log(U_1)}cos(2\pi U_2)$, where U_1 and U_2 are independent *Uni f* (0, 1) random variates; in fact, the quantity $\sqrt{-2\log(U_1)}sin(2\pi U_2)$ produces a second, independent $N(0,1)$ variate. This is the celebrated *Box-Muller* (1958) method. A $N(\mu, \sigma^2)$ variate can then be generated as $\mu + \sigma Z$.

- Double Exponential (or *Laplace*):$X \sim DE(\mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right)$$

$E(X) = \mu$ and $Var(X) = 2\sigma^2$.

Like the normal, this distribution is symmetric and unimodal, but has heavier tails and a somewhat different shape, being strictly concave up on both sides of μ.

- Logistic: $X \sim Logistic(\mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left[1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right]^2} .$$

  $E(X) = \mu$ and $Var(X) = (\pi^2/3)\sigma^2$.

  The logistic is another symmetric and unimodal distribution, more similar to the normal in appearance than the double exponential, but with even heavier tails.

- **t** (or *Student's t):* $X \sim t(\nu, \mu, \sigma^2)$, $x \in \Re$, $\nu > 0$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2, \nu) = \frac{\Gamma[(\nu+1)/2]}{\sigma\sqrt{\nu\pi}\Gamma(\nu/2)}\left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\nu+1)/2}$$

  $E(X) = \mu$ (if $\nu > 1$) and $Var(X) = \nu\sigma^2/(\nu - 2)$ (if $\nu > 2$). The parameter v is referred to as the *degrees of freedom* and is usually taken to be a positive integer, though the distribution is proper for any positive real number ν. The *t* is a common heavy-tailed (but still symmetric and unimodal) alternative to the normal distribution. The leading term in the pdf can be rewritten

$$\frac{\Gamma[(\nu+1)/2]}{\sigma\sqrt{\nu\pi}\Gamma(\nu/2)} = \frac{1}{\sigma\sqrt{\nu}B(\frac{1}{2}, \frac{\nu}{2})} ,$$

  where $B(.,.)$ again denotes the beta function.

- **Cauchy:** $X \sim Cau(\mu, \sigma^2) \equiv t(1, \mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{1}{\sigma\pi\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]} .$$

  $E(X)$ and $Var(X)$ do not exist, though μ is the median of this distribution. This special case of the *t* distribution has the heaviest possible tails and provides a wealth of counterexamples in probability theory.

- **Gamma:** $X \sim G(\alpha, \beta)$, $x > 0$, $\alpha > 0$, $\beta > 0$, and

$$p(x|\alpha, \beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} .$$

  $E(X) = \alpha\beta$ and $Var(X) = \alpha\beta^2$. Note also that if $X \sim G(\alpha, \beta)$, then $Y = cX \sim G(\alpha, c\beta)$.

  The gamma is a flexible family for continuous random variables defined on the positive real line. For $\alpha < 1$, the distribution has an infinite peak at 0 and is strictly decreasing. For $\alpha = 1$, the distribution intersects the vertical axis at $1/\beta$, and again is strictly decreasing (see the Exponential distribution below). For $\alpha > 1$, the distribution starts at the origin, increases for a time and then decreases; its appearance is roughly normal for large a. The gamma is the conjugate prior distribution for a Poisson rate parameter θ.

- **Exponential:** $X \sim Expo(\beta) \equiv G(1, \beta)$, $x > 0$, $\beta > 0$, and

$$p(x|\beta) = \frac{1}{\beta} e^{-x/\beta} .$$

$E(X) = \beta$ and $Var(X) = \beta^2$.

- **Chi-square:** $X \sim \chi^2(\nu) \equiv G(\nu/2, 2)$, $x > 0$, $\nu > 0$, and

$$p(x|\nu) = \frac{x^{\nu/2-1} e^{-x/2}}{\Gamma(\nu/2) 2^{\nu/2}} .$$

$E(X) = \nu$ and $Var(X) = 2\nu$. As with the t distribution, the parameter ν is referred to as the *degrees of freedom* and is usually taken to be a positive integer, though a proper distribution results for any positive real number ν.

- **Inverse Gamma:** $X \sim IG(\alpha, \beta)$, $x > 0$, $\alpha > 0$, $\beta > 0$, and

$$p(x|\alpha, \beta) = \frac{e^{-1/(\beta x)}}{\Gamma(\alpha) \beta^\alpha x^{\alpha+1}} .$$

$E(X) = 1/[\beta(\alpha-1)]$ (provided a > 1) and $Var(X) = 1/[\beta^2(\alpha-1)^2(\alpha-2)]$ (provided $\alpha > 2$). Note also that if $X \sim IG(\alpha, \beta)$, then $Y = cX \sim IG(\alpha, \beta/c)$.

A better name for the inverse gamma might be the *reciprocal gamma,* since $X = 1/Y$ where $Y \sim G(\alpha, \beta)$ (or in distributional shorthand, $IG(\alpha, \beta) \equiv 1/G(\alpha, \beta)$).

Despite its poorly behaved moments and somewhat odd, heavy-tailed appearance, the inverse gamma is very commonly used in Bayesian statistics as the conjugate prior for a variance parameter σ^2 arising in a normal likelihood function. Choosing α and β appropriately for such a prior can be aided by solving the above equations for $\mu \equiv E(X)$ and $\tau^2 \equiv Var(X)$ for α and β. This results in

$$\alpha = (\mu/\tau)^2 + 2 \quad \text{and} \quad \beta = \frac{1}{\mu\left[(\mu/\tau)^2 + 1\right]} .$$

Setting the prior mean and standard deviation both equal to μ (a reasonably vague specification) thus produces $\alpha = 3$ and $\beta = 1/(2\mu)$.

- Inverse Gaussian (or *Wald*): $X \sim InvGau(\mu, \lambda)$, $x > 0$, $\mu > 0$, $\lambda > 0$, and

$$p(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right]$$

$E(X) = \mu$ and $Var(X) = \mu^3/\lambda$.

The inverse Gaussian has densities that resemble those of the gamma distribution. Despite its somewhat formidable form, all its positive and negative moments exist.

A.2.2 Multivariate

- **Dirichlet:** $\mathbf{X} \sim D(\boldsymbol{\alpha})$, $\mathbf{X} = (x_1, \ldots, x_k)'$ where $0 \leq x_i \leq 1$ and $\sum_{i=1}^k x_i = 1$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ where $\alpha_i \geq 0$, and

$$p(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$$

  $E(X_i) = \alpha_i/\alpha_0$, $Var(X_i) = [\alpha_i(\alpha_0 - \alpha_i)]/[\alpha_0^2(\alpha_0 + 1)]$, and finally $Cov(X_i, X_j) = -(\alpha_i\alpha_j)/[\alpha_0^2(\alpha_0 + 1)]$, where $\alpha_0 \equiv \sum_{i=1}^k \alpha_i$.

  The Dirichlet is the multivariate generalization of the beta: note that for $k = 2$, $D(\boldsymbol{\alpha}) = Beta(\alpha_1, \alpha_2)$. Note that this is really only $(k - 1)$-dimensional distribution, since $x_k = 1 - \sum_{i=1}^{k-1} x_i$. The Dirichlet is the conjugate prior for the multinomial likelihood. It also forms the foundation for the *Dirichlet process prior,* the basis of most nonparametric Bayesian inference (see Section 2.5).

- **Multivariate Normal** (or *Multinormal*, or *Multivariate Gaussian*): $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{x} \in \Re^k$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ where $-\infty < \mu_i < \infty$, $\boldsymbol{\Sigma}$ a $k \times k$ positive definite matrix, and

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{k/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right] ,$$

  where $|\boldsymbol{\Sigma}^{-1}|$ denotes the determinant of $\boldsymbol{\Sigma}^{-1}$. $E(\mathbf{X}) = \boldsymbol{\mu}$ (i.e., $E(X_i) = \mu_i$ for all i), and $Var(X) = \boldsymbol{\Sigma}$ (i.e., $Var(X_i) = \sigma_{ii}$ and $Cov(X_i, X_j) = \sigma_{ij}$ where $\boldsymbol{\Sigma} = (\sigma_{ij})$). The multivariate normal forms the basis of the likelihood in most common linear models, and also serves as the conjugate prior for mean and regression parameters in such likelihoods.

  A $N_k(\mathbf{0}, \mathbf{I})$ variate $\mathbf{z}$ can be generated simply as a vector of k independent, univariate $N(0, 1)$ random variates. For general covariance matrices $\boldsymbol{\Sigma}$, we first factor the matrix as $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is a lower-triangular matrix (this is often called the *Cholesky factorization;* see for example Thisted, 1988, pp.81-83). A $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variate z can then be generated as $\boldsymbol{\mu} + \mathbf{L}\mathbf{z}$.

- **Multivariate t** (or *Multi-t):* $\mathbf{X} \sim t_k(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{x} \in \Re^k$, $\nu > 0$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ where $-\infty < \mu_i < \infty$, $\boldsymbol{\Sigma}$ a $k \times k$ positive definite matrix, and $p(\mathbf{x}|\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}\Gamma[(\nu + k)/2]}{(\nu\pi)^{k/2}\Gamma(\nu/2)} \left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\nu+k)/2}$$

$E(\mathbf{X}) = \boldsymbol{\mu}$ (if $\nu > 1$), and $Var(\mathbf{X}) = \nu\boldsymbol{\Sigma}/(\nu - 2)$ (if $\alpha > 2$). The multivariate *t* provides a heavy-tailed alternative to the multivariate normal while still accounting for correlation among the elements of **x** Typically, $\boldsymbol{\Sigma}$ is called a scale matrix, and is approximately equal to the variance matrix of X for large ν.

- **Wishart**: $\mathbf{V} \sim W(\boldsymbol{\Omega}, \nu)$, $\mathbf{V}$ a $k \times k$ symmetric and positive definite matrix, $\boldsymbol{\Omega}$ a $k \times k$ symmetric and positive definite parameter matrix , $\nu > 0$, and

$$p(\mathbf{V}|\nu, \boldsymbol{\Omega}) = c\frac{|\mathbf{V}|^{(\nu-k-1)/2}}{|\boldsymbol{\Omega}|^{\nu/2}} \exp\left[-\frac{1}{2} tr(\boldsymbol{\Omega}^{-1}\mathbf{V})\right] ,$$

provided the shape parameter (or "degrees of freedom") $\nu \geq k$. Here the proportionality constant c takes the awkward form

$$c = \left[2^{(\nu k)/2}\pi^{k(k-1)/4} \prod_{j=1}^{k} \Gamma\left(\frac{\nu + 1 - j}{2}\right)\right]^{-1} .$$

$E(V_{ij}) = \nu\Omega_{ij}$, $Var(V_{ij}) = \nu(\Omega_{ij}^2 + \Omega_{ii}\Omega_{jj})$, and finally $Cov(V_{ij}, V_{kl}) = \nu(\Omega_{ik}\Omega_{jl} + \Omega_{il}\Omega_{jk})$.

The Wishart is a multivariate generalization of the gamma, originally derived as the sampling distribution of the sum of squares and crossproducts matrix., $\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$, where $\mathbf{X}_i \overset{iid}{\sim} N_k(\boldsymbol{\mu}, \mathbf{V})$. In Bayesian analysis, just as the reciprocal of the gamma (the inverse gamma) is often used as the conjugate prior for a variance parameter σ^2 in a normal likelihood, the reciprocal of the Wishart (the *inverse Wishart*) is often used as the conjugate prior for a variance-covariance *matrix* $\boldsymbol{\Sigma}$ in a multivariate normal likelihood (see Example 5.6).

A random draw from a Wishart with ν an integer can be obtained using the idea in the preceding paragraph: If $\mathbf{x}_1, \ldots, \mathbf{x}_\nu$ are independent draws from a $N_k(\mathbf{0}, \boldsymbol{\Omega})$ distribution, then $\mathbf{V} = \sum_{i=1}^{\nu} \mathbf{x}_i\mathbf{x}_i'$ is a $W(\boldsymbol{\Omega}, \nu)$ random variate. Non-integral ν requires the general algorithm originally given by Odell and Feiveson (1966); see also the textbook treatment by Gentle (1998, p.107), or the more specifically Bayesian treatment by Gelfand et al. (1990).

APPENDIX B

Decision theory

B.1 Introduction

As discussed in Chapters 1, 3, and 4, the Bayesian approach with low-information priors strikes an effective balance between frequentist robustness and Bayesian efficiency. It occupies a middle ground between the two and, when properly tuned, can generate procedures with excellent frequentist properties, often "beating the frequentist approach at its own game." As an added benefit, the Bayesian approach more easily structures complicated models and inferential goals.

To make comparisons among approaches, we need a method of keeping score, commonly referred to as a loss *structure.* Loss structures may be explicit or implicit; here we discuss the structure adopted in a *decision-theoretic* framework. We outline the general decision framework, highlighting those features that suit the present purpose. We also provide examples of how the theory can help the practitioner. Textbooks such as Ferguson (1967) and DeGroot (1970) and the recent essay by Brown (2000) provide a more comprehensive treatment of the theory.

Setting up the general decision problem requires a prior distribution, a sampling distribution, a class of allowable actions and decision rules, and a loss function. Specific loss function forms correspond to point estimation, interval estimation, and hypothesis testing. We use the notation

$$\begin{aligned} \textit{prior distribution:} \quad & G(\theta),\ \theta \in \Theta \\ \textit{sampling distribution:} \quad & f(\mathbf{x}|\theta) \\ \textit{allowable actions:} \quad & a \in \mathcal{A} \\ \textit{decision rules:} \quad & d \in \mathcal{D} : \mathcal{X} \rightarrow \mathcal{A} \\ \textit{loss function:} \quad & l(\theta, a)\ . \end{aligned} \tag{B.1}$$

The loss function $l(\theta, a)$ computes the loss incurred when θ is the true state of nature and we take action a. Thus for point estimation of a parameter θ, we might use *squared error loss* (SEL),

$$l(\theta, a) = (\theta - a)^2$$

or *weighted squared error loss* (WSEL),

$$l(\theta, a) = w(\theta)(\theta - a)^2 ,$$

or *absolute error loss,*

$$l(\theta, a) = |\theta - a| ,$$

or, for discrete parameter spaces, 0-1 loss,

$$l(\theta, a) = \begin{cases} 0 & \text{if } \theta = a \\ 1 & \text{if } \theta \neq a \end{cases} . \tag{B.2}$$

The decision rule d maps the observed data $\mathbf{x}$ into an action a; in the case of point estimation, this action is a proposed value for θ (e.g., $d(x) = \bar{x}$

The Bayesian outlook on the problem of selecting a decision rule is as follows. In light of the data $\mathbf{x}$, our opinion as to the state of nature is summarized by the *posterior* distribution of θ, which is given by Bayes' Theorem as

$$dG(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)dG(\theta)}{m_G(\mathbf{x})} ,$$

where

$$m_G(\mathbf{x}) = \int f(\mathbf{x}|u)dG(u) .$$

The Bayes rule minimizes the *posterior risk,*

$$\rho(G, d(\mathbf{x})) = E_{\theta|\mathbf{x}}[l(\theta, d(\mathbf{x}))] = \int l(\theta, d(\mathbf{x}))dG(\theta|\mathbf{x}) . \tag{B.3}$$

Note that posterior risk is a single number regardless of the dimension of θ, so choosing d to minimize $\rho(G, d(\mathbf{x}))$ is well defined.

B.1.1 Risk and admissibility

In the frequentist approach, we have available neither a prior nor a posterior distribution. For a decision rule d, define its *frequentist risk* (or simply *risk*) for a given true value of θ as

$$R(\theta, d) = E_{\mathbf{x}|\theta}[l(\theta, d(\mathbf{x}))] = \int l(\theta, d(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x} , \tag{B.4}$$

the average loss, integrated over the distribution of X conditional on 0. Note that frequentist risk is a *function of* θ not a single number like the posterior risk (B.3).

Example B.1 In the interval estimation setting of Example 1.1, consider the loss function

$$l(\theta, a) = \begin{cases} 0 & \text{if } \theta \in \delta(\mathbf{x}) \\ 1 & \text{if } \theta \notin \delta(\mathbf{x}) \end{cases} .$$

Then the risk function is

$$\begin{aligned} R(8, d) &= E_{\mathbf{X}|\theta,\sigma^2}[l(0, \delta(\mathbf{X}))] \\ &= P_{\mathbf{X}|\theta,\sigma^2}[\theta \quad 6(X)] \\ &= .05 \, , \end{aligned}$$

which is constant over all possible values of θ and σ^2. ■

This controlled level of risk across all parameter values is one of the main selling points of frequentist confidence intervals. Indeed, plotting R versus θ for various candidate rules can be very informative. For example, if d_1 and d_2 are two rules such that

$$R(\theta, d_1) \ \# \ R(\theta, d_2) \quad \text{for all } \theta \, ,$$

with strict inequality for at least one value of θ, then under the given loss function we would never choose d_2, since its risk is never smaller that d_1's and can be larger. In such a case, the rule d_2 is said to be *inadmissible,* and *dominated* by d_1. Note that d_1 may itself be inadmissible, since there may be yet another rule which uniformly beats d_1. If no such rule exists, d_1 is called *admissible.*

Admissibility is a sensible and time-honored criterion for comparing decision rules, but its utility in practice is rather limited. While inadmissible rules often need not be considered further, for most problems there will be many admissible rules, creating the problem of which one to select. Moreover, admissibility is not a guarantee of sensible performance; the example in Subsection 4.2.1 provides an illustration. Therefore, additional criteria are required to select a frequentist rule. We consider the three most important such criteria in turn.

B.1.2 Unbiased rules

Unbiasedness is a popular method for reducing the number of candidate decision rules and allowing selection of a "best" rule within the reduced class. A decision rule $d(x)$ *is* unbiased if

$$E_{x|\theta}[l(\theta', d(x))] \geq E_{x|\theta}[l(\theta, d(x))] \text{ for all } \theta \text{ and } \theta' \, . \tag{B.5}$$

Under squared error loss, (B.5) is equivalent to requiring $E_{x|\theta}[d(x)] = \theta$. For interval estimation, unbiasedness requires that the interval have a greater chance of covering the true parameter than any other parameter. For hypothesis testing, unbiasedness is equivalent to the statistical power being greater under an alternative hypothesis than under the null.

Though unbiasedness has the attractive property of reducing the number of candidate rules and allowing the frequentist to find an optimal rule, it is often a very high price to pay - even for frequentist evaluations. Indeed, a principal thesis of this book is that a little bias can go a long way in improving frequentist performance. In Section B.2 below we provide the basic example based on squared error loss. Here we present three estimation examples of how unbiasedness can work against logical and effective inference.

Example B.2 : Estimating P(no events) Ferguson (1967) provides the following compelling example of a problem with unbiased estimates. Assume a Poisson distribution and consider the goal of estimating the probability that there will be no events in a time period of length $2t$ based on the observed number of events, X, in a time period of length t. Therefore, with λ the Poisson parameter we want to estimate $e^{-2t\lambda}$, based on X which is distributed Poisson$(t\lambda)$. The MLE of this probability is e^{-2X}, while the best (indeed, the only) unbiased estimate is $(-1)^X$. This latter result is derived by matching terms in two convergent power series.

This unbiased estimate is patently absurd; it is either +1 or -1 depending on whether an even or an odd number of events has occurred in the first t time units. Using almost any criterion other than unbiasedness, the MLE is preferred. For example, it will have a substantially smaller MSE. ■

Example B.3: Constrained parameter space. In many models the unbiased estimate can lie outside the allowable parameter space. A common example is the components of variance models (i.e., the Type II ANOVA model), where

$$Y_{ij} = \mu + b_i + \epsilon_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

where $\mu \in \Re$, $b_i \overset{iid}{\sim} N(0, \tau^2)$, and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. The unbiased estimate of τ^2, computed using the within and between mean squared errors, can sometimes take negative values (as can the MLE). Clearly, performance of these estimates is improved by constraining them to be nonnegative. ■

Example B.4: Dependence on the stopping rule. As with all frequentist evaluations, unbiasedness requires attention to the sampling plan and, if applicable, to the stopping rule used in collecting the data. Consider again Example 1.2, where now we wish to estimate the unknown success probability based on x observed successes in n trials. This information unambiguously determines the MLE as x/n, but the unbiased estimate depends on how the data were obtained. If n is fixed, then the MLE is unbiased. But if the sampling were "inverse" with data gathered until failures were obtained, then the number of observed successes, x, is negative binomial, and the unbiased estimate is $\frac{x}{x+r-1} = \frac{x}{n-1}$. If both n and x are random, it may be impossible to identify an unbiased estimate. ■

B.1.3 Bayes rules

While many frequentists would not even admit to the existence of a subjective prior distribution $G(\theta)$, a possibly attractive frequentist approach to choosing a best decision rule does depend on the formal use of a prior. For a prior G, define the *Bayes risk* as

$$r(G, d) = E_\theta E_{\mathbf{x}|\theta} l(\theta, d(\mathbf{x})) = E_\theta R(\theta, d) \ , \tag{B.6}$$

the expected frequentist risk with respect to the chosen prior G. Bayes risk is alternatively known as *empirical Bayes risk* because it averages over the variability in both θ and the data. Reversing the order of integration in (B.6), we obtain the alternate computational form

$$r(G, d) = E_{\mathbf{x}} E_{\theta|\mathbf{x}} l(\theta, d(\mathbf{x})) = E_{\mathbf{x}} \rho(G, d(\mathbf{x})) \ ,$$

the expected posterior risk with respect to the marginal distribution of the data $\mathbf{X}$. Since this is the posterior loss one expects *before* having seen the data, the Bayes risk is sometimes referred to as the *preposterior risk*. Thus $r(G, d)$ is directly relevant for Bayesian experimental design (see Section 4.5).

Since $r(G, d)$is a scalar quantity, we can choose the rule d_G that minimizes the Bayes risk, i.e.,

$$d_G(\mathbf{x}) = \arg\min_{d \in \mathcal{D}} r(G, d) \ . \tag{B.7}$$

This minimizing rule is called the *Bayes rule.* This name is somewhat confusing, since a subjective Bayesian would not average over the data (as $r(G, d)$ does) when choosing a decision rule, but instead choose one that minimized the posterior risk (B.3) given the observed data. Fortunately, it turns out that these two operations are virtually equivalent: under very broad conditions, minimizing the Bayes risk (B.6) is equivalent to minimizing the posterior risk (B.3) for all x such that $m_G(\mathbf{x}) > 0$. For further discussion of this result and related references, see Berger (1985, p.159).

Example B.5: Point estimation. Under SEL the Bayes rule is found by minimizing the posterior risk,

$$\rho(G, a) = \int (\theta - a)^2 g(\theta|\mathbf{x}) d\theta \ .$$

Taking the derivative with respect to a, we have

$$\frac{\partial}{\partial a}[\rho(G, a)] = \int 2(\theta - a)(-1) g(\theta|\mathbf{x}) d\theta \ .$$

Setting this expression equal to zero and solving for a, we obtain

$$a = \int \theta \, g(\theta|\mathbf{x}) d\theta = E(\theta|\mathbf{x}) \ ,$$

the posterior mean, as a solution. Since

$$\frac{\partial^2}{\partial a^2}[\rho(G, a)] = 2 \int g(\theta|\mathbf{x}) d\theta = 2 > 0 \ ,$$

the second derivative test implies our solution is indeed a minimum, confirming that the posterior mean is the Bayes estimate of θ under SEL.

We remark that under absolute error loss, the Bayes estimate is the posterior median, while under 0-1 loss, it is the posterior mode. ■

Example B.6: Interval estimation. Consider the loss function

$$l(\theta, a) = I_{\{\theta \notin a\}} + \text{c x volume(a)},$$

where *I* is the indicator function of the event in curly brackets and a is a subset of the parameter space Θ. Then the Bayes rule will be the region (or regions) for θ having highest posterior density, with c controlling the tradeoff between the volume of *a* and the posterior probability of coverage. Subsection 2.3.2 gives a careful description of Bayesian confidence intervals; for our present purpose we note (and Section 4.3 demonstrates) that under noninformative prior distributions, these regions produce confidence intervals with excellent frequentist coverage and modest volume. ■

Example B.7: Hypothesis testing. Consider the comparison of two simple hypotheses, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, so that we have a two-point parameter space $\Theta = \{\theta_0, \theta_1\}$. We use the "$0$–$l_i$" loss function

$$l(\theta, a) = al_0 + (1 - a)l_1,$$

where both l_0 and 1_1 are nonnegative, and a $\in$ {0, 1} gives the index of the accepted hypothesis. Then it can be shown that the Bayes rule with respect to a prior on $\pi = P(\theta = \theta_0)$ is a likelihood ratio test; in fact, many authors use this formulation to prove the Neyman-Pearson lemma. This model is easily generalized to a countable parameter space. ■

B.1.4 Minimax rules

A final (and very conservative) approach to choosing a best decision rule in a frequentist framework is to control the worst that can happen: that is, to minimize the maximum risk over the parameter space. Mathematically, we choose the rule $d*$ satisfying the relation

$$\sup_{\theta \in \Theta} R(\theta, d^*) = \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, d) \; .$$

This $d*$ may sacrifice a great deal for this control, since its risk may be unacceptably high in certain regions of the parameter space. The minimax rule is attempting to produce controlled risk behavior for all θ *and x,* but is making no evaluation of the relative likelihood of the θ values.

The Bayesian approach is helpful in finding minimax rules via a theorem which states that a Bayes rule (or limit of Bayes rules) that has constant risk over the parameter space is minimax. Generally, the minimax rule is not unique (indeed, in Chapter 3 we find more than one), and failure to find a constant risk Bayes rule does not imply that the absence of a minimax rule. Of course, as with all other decision rules, the minimax rule depends on the loss function. For the Gaussian sampling distribution, the sample mean is minimax (it is constant risk and the limit of Bayes rules). Finding minimax rules for the binomial and the exponential distributions have been left as exercises.

B.2 Procedure evaluation and other unifying concepts

B.2.1 Mean squared error

To see the connection between the Bayesian and frequentist approaches, consider again the problem of estimating a parameter under squared error loss (SEL). With the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, let $d(\mathbf{X})$ be our estimate of θ. Its frequentist risk is

$$E_{\mathbf{x}|\theta}[l(\theta, d(\mathbf{x}))] = E_{\mathbf{x}|\theta}[(\theta - d(\mathbf{x}))^2] \equiv MSE_d(\theta) \ ,$$

the *mean squared error* of d given the true θ. Its posterior risk with respect to some prior distribution G is

$$E_{\theta|\mathbf{x}}[l(\theta, d(\mathbf{x}))] = E_{\theta|\mathbf{x}}[(\theta - d(\mathbf{x}))^2] \equiv MSE_{d,G}(\mathbf{x}) \ ,$$

and its preposterior risk is

$$E_{\theta,\mathbf{x}}[l(\theta, d(\mathbf{x}))] = E_{\theta,\mathbf{x}}[(\theta - d(\mathbf{x}))^2] \equiv MSE_{d,G} \ ,$$

which equals $E_\theta[MSE_d(\theta)]$ and $E_{\mathbf{x}}[MSE_{d,G}(\mathbf{x})]$. We have not labeled d as "frequentist" or "Bayesian;" it is simply a function of the data. Its properties can be evaluated using any frequentist or Bayesian criteria.

B.2.2 The variance-bias tradeoff

We now indicate how choosing estimators having minimum MSE requires a tradeoff between reducing variance and bias. Assume that conditional on θ, the data are i.i.d. with mean θ and variance $\sigma^2(\theta)$. Let $d(\mathbf{x})$ be the sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and consider estimators of θ of the form

$$d_c(\mathbf{x}) = cX \ . \qquad \text{(B.8)}$$

For 0 < c < 1, this estimator shrinks the sample mean toward 0 and has the following risks:

$$\begin{aligned}
\text{frequentist:} \quad & c^2\frac{\sigma^2(\theta)}{n} + (1-c)^2\theta^2 \\
\text{posterior:} \quad & Var(\theta|\mathbf{x}) + [c\bar{x} - E(\theta|\mathbf{x})]^2, \\
\text{preposterior (Bayes):} \quad & c^2\frac{E[\sigma^2(\theta)]}{n} + (1-c)^2\{Var(\theta) + [E(\theta)]^2\},
\end{aligned}$$

where in the last expression the expectations and variances are with respect to the prior distribution G. Notice that all three expressions are of the form "variance + bias squared."

Focusing on the frequentist risk, notice that the first term decreases like $\frac{1}{n}$ and the second term is constant in n. If $\theta = 0$ or is close to 0, then a c near 0 will produce small risk. More generally, for relatively small n, it will be advantageous to set c to a value betweeen 0 and 1 (reducing the variance at the expense of increasing the bias), but, as n increases, c should converge to 1 (for large n, bias reduction is most important). This phenomenon underlies virtually all of statistical modeling, where the standard way to reduce bias is to add terms (degrees of freedom, hence more variability) to the model.

Taking c = 1 produces the usual unbiased estimate with MSE equal to the sampling variance. Comparing c = 1 to the general case shows that if

$$\frac{n\theta^2 - \sigma^2(\theta)}{n\theta^2 + \sigma^2(\theta)} < c \leq 1,$$

then it will be advantageous to use the biased estimator. This relation can be turned around to find the interval of θs around 0 for which the estimator beats $\bar{X}$. This interval has endpoints that shrink to 0 like $\frac{1}{\sqrt{n}}$.

Turning to the Bayesian posterior risk, the first term also decreases like $\frac{1}{n}$, and the Bayes rule (not restricted to the current form) will be $E(\theta|\mathbf{x})$. The c that minimizes the preposterior risk is

$$c = c_n = \frac{n\{Var(\theta) + [E(\theta)]^2\}}{n\{Var(\theta) + [E(\theta)]^2\} + E[\sigma^2(\theta)]} .$$

As with the frequentist risk, for relatively small n it is advantageous to use $c \neq 1$ and have c_n increase to 1 as n increases.

Similar results to the foregoing hold when an estimator (B.8) is augmented by an additive offset (i.e., $d_{a,c}(\mathbf{x}) = a + c\bar{X}$). We have considered SEL and MSE for analytic simplicity; qualitatively similar conclusions and relations hold for a broad class of convex loss functions.

Shrinkage estimators of the form (B.8) arise naturally in the Bayesian context. The foregoing discussion indicates that they can have attractive frequentist properties (i.e., lower MSE than the standard estimator, $\bar{X}$). Under squared error loss, one wants to strike a tradeoff between variance and bias, and rules that effectively do so will have good properties.

B.3 Other loss functions

We now evaluate the Bayes rules for three loss functions that show the effectiveness of Bayesian structuring and how intuition must be aided by the rules of probability. Chapter 4 and Section 7.1 present additional examples of the importance of Bayesian structuring.

B.3.1 Generalized absolute loss

Consider the loss function for $0 \leq p \leq 1$,

$$\begin{array}{rcl} p|\theta - a| & \text{if} & a < \theta \\ (1-p)|\theta - a| & \text{if} & a \geq \theta \ . \end{array}$$

It is straightforward to show that the optimal estimate is the p^{th} percentile of the posterior distribution. Therefore, p = .5 gives the posterior median. Other values of p are of more than theoretical interest, as illustrated below in Problem 3.

B.3.2 Testing with a distance penalty

Consider a hypothesis testing situation wherein there is a penalty for making an incorrect decision and, if the null hypothesis is rejected, an additional penalty that depends on the distance of the parameter from its null hypothesis value. The loss function

$$l(\theta,\ a) = \begin{cases} 0 & , \quad a = 0, \theta < 0 \text{ or } a = 1, \theta > 0 \\ \theta^2 & , \quad a = 0, \theta > 0 \text{ or } a = 1, \theta < 0 \end{cases}$$

computes this score. Assume the posterior distribution is $N(\mu, \tau^2)$. Then

$$R(\theta, a = 1) - R(\theta, a = 0) = \tau^2[1 - 2\Phi(\mu/\tau)] + \frac{\mu}{\tau}\phi(\mu/\tau) \ ,$$

where the first term corresponds to 0-1 loss, while the second term corresponds to the distance adjustment. If $\mu > 0$, 0-1 loss implies $a = 1$, but with the distance adjustment, if $\tau < 1$, then for small $\mu > 0$, $a = 0$. This decision rule is non-intuitive, but best.

B.3.3 A threshold loss function

Through its Small Area Income and Poverty Estimates (SAIPE) project, the Census Bureau is improving estimated poverty counts and rates used in allocating Title I funds (National Research Council, 2000). Allocations must satisfy a "hold-harmless" condition (an area-specific limit on the reduction in funding from one year to the next) and a threshold rule (e.g., concentration grants kick in when estimated poverty exceeds 15%). ACS (2000) demonstrates that statistical uncertainty induces unintended consequences of a hold-harmless condition, because allocations cannot appropriately adjust to changing poverty estimates. Similarly, statistical uncertainty induces unintended consequences of a threshold rule. Unintended consequences can be ameliorated by optimizing loss functions that reflect societal goals associated with hold-harmless, threshold, and other constraints. However, amelioration will be limited and societal goals may be better met through

condition	eligible for concentration funds?	loss
$\theta \geq T$	yes	$(\theta - a)^2$
$\theta < T$	no	a^2

Table B.1 *A threshold loss function.*

replacing these constraints by ones designed to operate in the context of statistical uncertainty.

Example B.8 To show the quite surprising influence of thresholds on optimal allocation numbers, we consider a greatly simplified, mathematically tractable example. Let θ be the true poverty rate for a single area (e.g., a county), a be the amount allocated to the area per child in the population base, and $\mathbf{Y}$ denote all data. For a threshold T, consider the societal loss function in Table B.8. The optimal per-child allocation value is then

$$a_T(\mathbf{Y}) = E(\theta \mid \theta \geq T, \mathbf{Y}) \times \mathrm{pr}(\theta \geq T \mid \mathbf{Y}) .$$

Though the allocation formula as a function of the true poverty rate (0) has a threshold, the optimal allocation value has no threshold. Furthermore, for $T > 0$, $a_T(\mathbf{Y})$ is not the center of the the posterior distribution (neither the mean, median, nor mode). It is computed by multiplying the posterior mean (conditional on $\theta > T$) by a posterior tail area. This computation always produces an allocation value smaller than the posterior mean; for example, if the posterior distribution is exponential with mean μ, $a_T = \mu(1 + r)e^{-r}$ where $r = T/\mu$. But this reduction is compensated by the lack of a threshold in a_T. It is hard to imagine coming up with an procedure in this setting without Bayesian structuring.

Actual societal loss should include multiplication by population size, incorporate a dollar cap on the total allocation over all candidate areas, and "keep score" over multiple years. Absolute error should replace squared-error in comparing a to θ. ■

B.4 Multiplicity

There are several opportunities for taking advantage of multiple analyses, including controlling error rates, multiple outcomes, multiple treatments, interim analyses, repeated outcomes, subgroups, and multiple studies. Each of these resides in a domain for control of statistical properties, including control for a single analysis, a set of endpoints, a single study, a collection of studies, or even an entire research career! It is commonly thought that the Bayesian approach pays no attention to multiplicity. For example, as

α	$Z_{\alpha/2}$	$Z_{\alpha_{bf}/2}$	$\sqrt{-2\log\alpha}$
.01	2.57	2.81	3.03
.02	2.33	2.57	2.80
.05	1.96	2.24	2.45
.10	1.65	1.95	2.15

Table B.2 *Comparison of upper univariate confindence limits, multiple comparisons setting with* $K = 2$.

shown in Example 1.2, the posterior distribution does not depend on a data monitoring plan or other similar aspects of an experimental design. However, a properly constructed prior distribution and loss function can acknowledge and control for multiplicity.

Example B.9 When a single parameter is of interest in a K-variate parameter structure, the Bayesian formulation automaticallly acknowledges multiplicity. For the $K = 2$ case, consider producing a confidence interval for θ_1 when: $\boldsymbol{\theta} = (\theta_1, \theta_2) \sim N_2(\mathbf{0}, I_2)$. The (1 - α HPD region is

$$|| \boldsymbol{\theta} ||^2 \leq -2\log(\alpha) .$$

Projecting this region to the first coordinate gives

$$-\sqrt{-2\log(\alpha)} \leq \theta_1 \leq \sqrt{-2\log(\alpha)} .$$

Table B.2 shows that for $K = 2$ the Bayesian HPD interval is longer than either the unadjusted or the Bonferroni adjusted intervals (α_{bf} is the Bonferroni non-coverage probability). That is, the Bayesian interval is actually more conservative! However, as K increases, the Bayesian HPD interval becomes narrower than the Bonferroni interval. ■

B.5 Multiple testing

B.5.1 Additive loss

Under additive, component-specific loss, Bayesian inference with components that are independent a priori separately optimizes inference for each component with no accounting for the number of comparisons. However, use of a hyperprior Bayes (or empirical Bayes) approach links the components, since the posterior "borrows information" among components. The *k*-ratio *t*-test is a notable example. With F denoting the F-test for a one-way ANOVA,

$$\begin{aligned} F &= (\hat{\sigma}^2 + K\hat{\tau}^2)/\hat{\sigma}^2 \\ (1 - \hat{B}) &= (F - 1)/F \end{aligned}$$

$$\text{and } Z_{12} = \left(\frac{F-1}{F}\right)^{1/2} \frac{Y_1 - Y_2}{\sqrt{2}\sigma}$$

The magnitude of F adjusts the test statistic. For large K, under the global null hypothesis $H_0 : P(\text{all } Z_{ij} = 0) \geq 0.5$, the rejection rate is much smaller than 0.5. Note that the procedure depends on the estimated prior mean $(\hat{\mu})$ and shrinkage $(\hat{B})$ and the number of candidate coordinates can influence these values. If the candidate coordinates indicate high heterogeneity, then $\hat{B}$ is small as is the Bayesian advantage. Therefore, it is important to consider what components to include in an analysis to control heterogeneity while retaining the opportunity for discovery.

B.5.2 Non-additive loss

If one is concerned about multiplicity, the loss function should reflect this concern. Consider a testing problem with a loss function that penalizes for individual coordinate errors and adds an extra penalty for making two errors:

$$\begin{aligned}
\text{parameters:} \quad & \theta_1, \theta_2 \in \{0, 1\} \\
\text{decisions:} \quad & a_1, a_2 \in \{0, 1\} \\
l(a, \boldsymbol{\theta}) \quad = \quad & a_1(1-\theta_1)\ell_0 + (1-a_1)\theta_1\ell_1 \\
& + a_2(1-\theta_2)\ell_0 + (1-a_2)\theta_2\ell_1 \\
& + \gamma(1-\theta_1)(1-\theta_2)a_1a_2.
\end{aligned}$$

With $\ell_0 = \ell_1 = 1$ and $G_{ij} = P(\theta_1 = i, \theta_2 =$ j data), a straightforward computation shows that the risk is

$$a_1(1 - 2G_{1+}) + a_2(1 - 2G_{+1}) + \gamma a_1 a_2 G_{00} + \{G_{1+} + G_{+1}\}\ .$$

The four possible values of (a_1, a_2) then produce the risks in Table B.3, and the the Bayes decision rule (the (a_1, a_2) that minimize risk) is given in Table B.4. Note that in order to declare the second component "$a_2 = 1$," we require more compelling evidence than if $\gamma = 0$.

a_1	a_2	risk $- \{G_{1+} + G_{+1}\}$
0	0	0
1	0	$1 - 2G_{1+}$
0	1	$1 - 2G_{+1}$
1	1	$(1 - 2G_{1+}) + (1 - 2G_{1+}) + \gamma G_{00}$

Table B.3 *Risks for the possible decision rules.*

condition	optimal rule
$G_{1+} \leq .5, G_{+1} \leq .5$	$a_1 = 0, a_2 = 0$
$G_{1+} \leq .5, G_{+1} > .5$	$a_1 = 0, a_2 = 1$
$G_{1+} > .5, G_{+1} \leq .5$	$a_1 = 1, a_2 = 0$
$G_{1+} > G_{+1} > .5$	$a_1 = 1,\ a_2 = \begin{cases} 0, & \text{if } (2G_{+1} - 1) < \gamma G_{00} \\ 1, & \text{if } (2G_{+1} - 1) \geq \gamma G_{00} \end{cases}$

Table B.4 *Optimal decision rule for non-additive loss.*

Discussion

Statistical decision rules can be generated by any philosophy under any set of assumptions. They can then be evaluated by any criteria - even those arising from an utterly different philosophy. We contend (and much of this book shows) that the Bayesian approach is an excellent "procedure generator," even if one's evaluation criteria are frequentist. This somewhat agnostic view considers the parameters of the prior, or perhaps the entire prior, as "tuning parameters" that can be used to produce a decision rule with broad validity. Of course, no approach automatically produces broadly valid inferences, even in the context of the Bayesian models. A procedure generated assuming a cocksure prior that turns out to be far from the truth will perform poorly in both the Bayesian and frequentist senses.

B.6 Exercises

1. Show that the Bayes estimate of a parameter θ under weighted squared error loss, $l(\theta, a) = w(\theta)(\theta - a)^2$, is given by

$$d_\pi(\mathbf{x}) = \frac{E[\theta w(\theta)|\mathbf{x}]}{E[w(\theta)|\mathbf{x}]} .$$

2. Suppose that the posterior distribution of θ, $p(\theta|\mathbf{x})$, is discrete with support points $\{\theta_1, \theta_2, \ldots\}$. Show that the Bayes rule under 0-1 loss (B.2) is the posterior mode.

3. Consider the following loss function:

$$l(\theta, a) = \begin{cases} p|\theta - a|, & \theta > a \\ (1-p)|\theta - a|, & \theta \leq a \end{cases}$$

Show that the Bayes estimate is the 100 x p^{th} percentile of the posterior distribution.

(Note: The posterior median is the Bayes estimate for p = .5. This loss function does have practical application for other values of p. In many organizations, employees are allowed to set aside pre-tax dollars for health care expenses. Federal rules require that one cannot get back funds not used in a 12-month period, so one should not put away too much. However, putting away too little forces some health care expenses to come from post-tax dollars. If losses are linear in dollars (or other currency), then this loss function applies with p equal to the individual's marginal tax rate.)

4. For the binomial distribution with n trials, find the minimax rule for squared error loss and normalized squared error loss, $L(\theta, a) = \frac{(\theta - a)^2}{\theta(1-\theta)}$. (*Hint:* The rules are Bayes rules or limits of Bayes rules.)

   For each of these estimates, compute the frequentist risk (the expected loss with respect to $f(x|\theta)$), and plot this risk as a function of θ (the *risk plot).* Use these risk plots to choose and justify an estimator when θ *is*

   (a) the head probability of a randomly selected coin;
   (b) the failure probability for high-quality computer chips;
   (c) the probability that there is life on other planets.

5. For the exponential distribution based on a sample of size n, find the minimax rule for squared error loss.

6. Find the frequentist risk under squared error loss (i.e., the MSE) for the estimator $d_{a,c}(\mathbf{x}) = a + cX$. What true values of θ favor choosing a small value of c (high degree of shrinkage)? Does your answer change if n is large?

7. Let X be a random variable with mean μ and variance σ^2. You want to estimate μ under SEL, and propose an estimate of the form $(1 - b)X$.

   (a) Find b^*, the b that minimizes the MSE.
   (*Hint:* Use the "variance + (bias)^2" representation of MSE.)
   (b) Discuss the dependence of b^* on μ and σ^2 and its implications on the role of shrinkage in estimation.

8. Consider a linear regression having true model of the form $Y_i = \alpha + \beta x_i + \epsilon_i$, where the ϵ_i are i.i.d. with mean 0 and variance σ^2. Suppose you have a sample of size n, use least-squares estimates (LSEs), and want to minimize the *prediction error:*

$$\text{PE} = E[Y_{new} - \hat{Y}(X_{new})^2] .$$

   Here, $Y(x)$ is the prediction based on the estimated regression equation and X_{new} is randomly selected from a distribution with mass $\frac{1}{n}$ at each of the $x_1, \ldots, x_n$.

(a) Assume that $\sum_i x_i = 0$ and show that even though you may *know* that $\beta \neq 0$, if $\sum_i x_i^2$ is sufficiently small, it is better to force $\hat{\beta} = 0$ than to use the LSE for it.
(Note: This result underlies the use of the C_p statistic and PRESS residuals in selecting regression variables; see the documentation for SAS Proc REG.)

(b) *(More difficult)* As in the previous exercise, assume you plan to use a prediction of the form $\hat{Y}(x) = \hat{\alpha} + (1-b)\hat{\beta}x$. Find $b*$, the optimal value of b.

9. In an errors-in-variables simple regression model, the least squares estimate of the regression slope (β) is biased toward 0, an example of *attenuation*. Specifically, if the true regression (through the origin) is $Y = x\beta + \epsilon$, but Y is regressed on X, with $X = x + \delta$, then the least squares estimate $(\hat{\beta})$ has expectation: $E[\hat{\beta}] \approx \rho\beta$, with $\rho = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\delta^2} \leq 1$.

If ρ is known or well-estimated, one can correct for attenuation and produce an unbiased estimate by using $\hat{\beta}/\rho$ to estimate β. However, this estimate may have poor MSE properties. To minimize MSE, consider an estimator of the form: $\hat{\beta}_c = c(\hat{\beta}/\rho)$.

(a) Let $\sigma^2 = Var[\hat{\beta}]$ and find the c that minimizes MSE.

(b) Discuss the solution's implications on the variance/bias tradeoff and the role of shrinkage in estimation.

10. Compute the HPD confidence interval in Example B.9 for a general K, and compare its length to the unadjusted and Bonferrom adjusted intervals.

APPENDIX C

Software guide

Many Bayesians argue that the continuing emphasis on classical rather than Bayesian methods in actual practice is less a function of philosophical disagreement than of the lack of reliable, widely-available software. While the largest commercial statistical software packages by and large do not provide Bayesian procedures, a large number of hierarchical modeling and empirical Bayes programs are commercially available. Moreoever, the number of fully Bayesian packages continues to burgeon, with many available at little or no cost. This appendix provides a brief guide to many of these software packages, with information on how to obtain some of the noncommercial products.

We caution the reader at the outset that we make no claim to provide full coverage, and that guides like this one go out of date very quickly. Few of the packages listed in the exhaustive review of Bayesian software by Goel (1988) remain in common use today. Packages evolve over time; some are expanded and combined with other packages, while others disappear altogether. Developers' addresses sometimes change, or primary responsibility for a package passes to a new person or group (e.g., when a successful noncommercial package makes the jump to commercial status, as in the case of *S* and *S-plus*).

Many of the programs listed below are freely and easily available electronically via the World Wide Web (WWW, or simply "the web"). Many of these are contained in *StatLib,* a system for the electronic distribution of software, datasets, and other information. Supported by the Department of Statistics at Carnegie Mellon University, *StatLib* can be accessed over the web at

`http://lib.stat.cmu.edu/`

The reader may also wish to consult the first author's webpage,

`http://www.biostat.umn.edu/~brad`

for more recently updated information on available software, as well as Bayes and EB inference in general. Finally, the recent overview by Berger (2000) includes an appendix with additional Bayesian software websites.

C.1 Prior elicitation

While computation of posterior and predictive summaries is the main computing task associated with applied Bayes or empirical Bayes analysis, it is certainly not the first. Prior distributions for the model parameters must be determined, and in many cases this involves the elicitation of prior opinion from decision makers or other subject-matter experts. Subsection 2.2.1 showed that this can be an involved process, and a number of computer programs for elicitation have been described in the literature.

Linear regression models

Perhaps the first such program of this type was reported by Kadane et al. (1980). These authors developed an interactive program for eliciting the hyperparameters of a conjugate prior distribution for the standard regression model (i.e., one having linear mean structure and normally distributed errors). Implemented on an experimental computer system called *TROLL* at the Massachusetts Institute of Technology, the program and its platform look fairly primitive by today's standards. However, the work showed that such programs were feasible, and contributed the important idea of eliciting prior opinion on potentially observable quantities (such as predicted values of the response variable) instead of the regression parameters themselves, about which the elicitee may have little direct feeling.

An updated Fortran version of the above approach is described in Kadane and Wolfson (1996), and freely available from *SiaiLib* at

http://lib.stat.cmu.edu/general/elicit-normlin/

Kadane, Chan, and Wolfson (1996) describe an extension of these methods to the case of AR(1) error structure. See also

http://lib.stat.cmu.edu/general/elicit-diric/

for similar code for eliciting a prior on the concentration parameter in a Dirichlet process prior. Finally, elicitation programs for several Bayesian models are available on Prof. Wolfson's software page,

http://statweb.byu.edu/ljw/elicitation/

Use of dynamic graphics

Dr. Russell Almond developed a program called ElToY for eliciting Bayesian conjugate prior distributions through constraint-based dynamic graphics. Written in XLISP-STAT, the program does one parameter elicitation for univariate conjugate priors through dynamic graphics. It contains a tool for interactively displaying a univariate distribution, as well as a demonstration of the central limit theorem. Copies are available over the web at

http://lib.stat.cmu.edu/xlispstat/eltoy/

C.2 Random effects models/ Empirical Bayes analysis

SAS **Proc MIXED**

Proc MIXED is a SAS procedure for analyzing linear models with both fixed and random effects. As we have seen, these models are particularly useful for the analysis of repeated measures data from clinical trials and cohort studies. The basic form dealt with is exactly that of equation (7.10), namely,

$$Y_i = \mathbf{X}_i \alpha + \mathbf{W}_i \beta_i + \epsilon_i \ , \ i = 1, \ldots, n,$$

where a is the vector of fixed effects and the are vectors of subject-specific random effects, assumed to be independent and normally distributed with mean vector **0** and covariance matrix The are also assumed independent (of each other and of the and identically distributed as normals with mean **0** and covariance matrix Thus, marginally we have that

$$Var(Y_i) = \mathbf{W}_i \mathbf{V} \mathbf{W}_i^{'} + \boldsymbol{\Sigma} \ ,$$

so that even if $\boldsymbol{\Sigma}$ is taken as $\sigma_i^2 \mathbf{I}$ (as is customary), the components of Y_i are still marginally correlated.

Proc MIXED essentially works with the marginal model above, obtaining point estimates of the model parameters by maximum likelihood, implemented via the Newton-Raphson algorithm. Individual level point and interval estimates for the parameters $\hat{\beta}_i$ and predicted values $\hat{Y}_i$ are also available. While this approach is essentially classical, if one thinks of the random effects distribution as a prior, it is the same as empirical Bayes. **Proc MIXED** addresses the problem of undercoverage of "naive" individual level confidence intervals (see Section 3.5) by using t-distributions with appropriately selected degrees of freedom.

Several forms for the covariance matrices **V** and $\boldsymbol{\Sigma}$ are available, ranging from a constant times the identity matrix (conditional independence) to completely unstructured. Popular intermediate choices include compound symmetry (constant correlation; equivalent to a random intercept) and various time-series structures where correlation depends on the temporal or spatial distance between observations. The package provides the data analyst with a rich array of modeling options for the mean vector, the random effects and the "time series" structures.

Another SAS procedure, **NLMIXED**, fits nonlinear mixed models; that is, models in which both fixed and random effects enter nonlinearly. These models have a wide variety of applications, two of the most common being pharmacokinetics and overdispersed binomial data. **Proc NLMIXED** enables you to specify a conditional distribution for your data (given the random effects) having either a standard form (normal, binomial, Poisson) or a general distribution that you code using SAS programming statements.

Proc NLMIXED fits nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effects. Different integral approximations are available, the principal ones being adaptive Gaussian quadrature and a first-order Taylor series approximation. A variety of alternative optimization techniques are available to carry out the maximization; the default is a dual quasi-Newton algorithm.

Proc NLMIXED is currently limited to only two-level models, and is also limited by its quadrature approach. However, two SAS macros are available to accommodate a much wider class of models. The first, **GLIMMIX,** allows fairly arbitrary generalized linear models, including repeated measures for binary or count data, and Poisson or logistic regression with random effects. It does this by iteratively applying **Proc MIXED** to a linear approximation to the model (see Wolfinger and O'Connell, 1993). The second, **NLINMIX,** is specifically set up to handle nonlinear random coefficient models. This macro works by iteratively calling **Proc MIXED** following an initial call to **Proc NLIN**. Wolfinger (1993) shows that this is mathematically equivalent to the second-order approximate method of Lindstrom and Bates (1990), which expands the nonlinear function about random effects parameters set equal to their current empirical best linear unbiased predictor. The macros are available at

http://www.sas.com/techsup/download/stat/

in the files beginning with "glmm" and "nlmm" prefixes.

Finally, we point out one fully Bayesian capability of **Proc MIXED**, first documented in SAS Release 6.11. Independent pseudo-random draws from the joint posterior distribution of the model parameters are available in the case of what are often referred to as *variance components models.* Searle, Casella, and McCulloch (1992) provide a comprehensive review of these models; an outstanding Bayesian treatment is given in Section 5.2 of Box and Tiao (1973). More specifically, Wolfinger and Rosner (1996) illustrate this capability for the hierarchical model

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma x_{ijkl} + u_{ijk} + \epsilon_{ijkl} \ ,$$

where the u_{ijk} and ϵ_{ijkl} are both independent, mean-zero normal errors with variances σ_u^2 and σ_e^2, respectively. Using the **PRIOR** statement to specify the prior on these two parameters (flat or Jeffreys), **Proc MIXED** then specifies flat priors for the remaining (fixed) effects, and integrates them out of the model in closed form. After improving the symmetry of the remaining bivariate posterior surface via linear transformation. the program generates posterior values for σ_u^2 and σ_e^2 via an independence chain algorithm which uses a product of inverse gamma densities as its base sampling function. Since the conditional posterior of the remaining parameters given σ_u^2 and σ_e^2 is multivariate normal, the full posterior sample is easily completed. This sample can be output to a SAS dataset, and subsequently analyzed using **Proc UNIVARIATE**, **Proc CAPABILITY**, **Proc KDE2D** (for two-dimensional kernel density estimates and contour plots), or the exploratory data analytic SAS/INSIGHT software. For more details and examples, refer to Wolfinger and Kass (2000). General information and updates about the preceding SAS software is available at the company's webpage at

http://www.sas.com/rnd/app/Home.html

Other commercial packages

Several other software packages for fitting random effects and other mixed models are available commercially. Kreft et al. (1994) provide a review of five of these packages: **BMDP-5V**, **GENMOD**, **HLM**, **ML3**, and **VARCL**. These authors comment specifically on how these five packages compare with respect to ease of use, documentation, error handling, execution speed, and output accuracy and readability. Zhou et al. (1999) offer a similar comparative review of four packages that can handle discrete response data *(generalized* linear mixed models): SAS **Proc MIXED** (via the old **Glimmix** macro), **MLwiN, HLM, and VARCL**. Here, we merely sketch some of the features of these packages.

Statistical and biostatistical readers will likely be most familiar with **BMDP-5V**, a procedure in the BMDP biomedical data analysis package initiated in the 1960s at the University of California. Like **Proc MIXED**, **5V** is intended for the analysis of repeated measures data, and contains many of the same modeling features, such as a generous collection of options for the forms of the variance matrices $\mathbf{V}$ and $\boldsymbol{\Sigma}$. It also allows several unusual designs. such as unbalanced or partially missing data (imputed via the EM algorithm), and time-varying covariates (see Schluchter, 1988). However, at present it has no capability for fully Bayesian posterior analysis.

The remaining packages are targeted to specific groups of practitioners. **GENMOD** is popular among demographers, the group for whom it was originally developed. **HLM** is widely used by social scientists and educational researchers in the U.S. (where it was developed), while **ML3** is a similar program more popular in the U.K. (where it was developed by Prof. Harvey Goldstein and colleagues; c.f. Goldstein, 1995). **VARCL**, also developed in the U.K., is also used by educational researchers. These PC-based programs tend to be smaller and less expensive than SAS or BMDP.

ML3 has been replaced by **MLn**, which expands the program's modeling capability from three to an arbitrary number of levels, and by **MLwiN**, the Windows successor to **MLn**. The program is distributed by the Multilevel Models Project at the Institute of Education, University of London. Their web homepage can be accessed at

http://multilevel.ioe.ac.uk/

MLwiN now offers the capability for fully Bayesian analysis in a limited class of models under diffuse priors using MCMC methods. Specifically, Gibbs sampling is used with normal responses, and Metropolis-Hastings sampling for normal and binary proportion responses; capability for general binomial and Poisson response models will be forthcoming shortly. Sample trace plots, kernel density estimates, and several convergence diagnostics are also implemented. For information, see

http://www.ioe.ac.uk/mlwin/mcmc.html/

MLwiN, **HLM**, and **VARCL** are also distributed by ProGamma, an interuniversity expertise center which develops and distributes a very wide range of computer applications for the social and behavioral sciences. Information on obtaining these and many other popular programs is available at

http://www.gamma.rug.ni/

Noncommercial packages

Noncommercial packages (freeware available from their developers or distributed over computer networks) continue to be developed to fit hierarchical and empirical Bayes models. In the case of likelihood analysis, one of the best documented and easiest to use alternatives is **nlme**, a collection of *S* functions, methods, and classes for the analysis of linear and nonlinear mixed-effects models written by Drs. Douglas Bates, Jose Pinheiro, and Mary Lindstrom. See

http://nlme.stat.wisc.edu/

for a full description. Both the *S* and the *C* source code are freely available from this website. The commercial vendor MathSoft does ship this code with various versions of *S-plus*, but the most up-to-date code is always that from the above web site. For more on *S-plus* itself, see

http://www.insightful.com/products/default.asp

and the book by Pinheiro and Bates (2000) for the connection to **nlme**.

A similar approach is provided by the **MIXOR** and **MIXREG** programs by Drs. Donald Hedeker and Robert D. Gibbons. These programs include mixed-effects linear regression, mixed-effects logistic regression for nominal or ordinal outcomes, mixed-effects probit regression for ordinal outcomes, mixed-effects Poisson regression, and mixed-effects grouped-time survival analysis. For more information, see

http://www.uic.edu/~hedeker/mix.html

Perhaps the most noteworthy current program for empirical Bayes analysis is **PRIMM**, an acronym for *Poisson Regression Interactive Multilevel Model*. **PRIMM** is an *S-plus* function for analyzing a three-stage Poisson regression model under the usual log link function. The program uses a relatively vague hyperprior at the third model stage so that the resulting estimates of the regression and individual-specific parameters generally enjoy good frequentist and Bayesian properties, similar to our perspective in Chapter 4. Some aspects of the hyperprior may be input by the user or simply set equal to data-based default values, corresponding to a Bayes or EB analysis, respectively. In the latter case, an extension of Morris' EB interval adjustment method presented in Subsection 3.5.1 is used to avoid the undercoverage associated with naive EB confidence intervals.

PRIMM's computational method combines standard maximization routines with clever analytic approximations to obtain results much more quickly (though somewhat less accurately) than could be achieved via MCMC methods. Christiansen and Morris (1997) give a full description of the model. prior distributions, computational approach, and several supporting data analyses and simulation studies. The *S* code is available on *StatLib* at

http://lib.stat.cmu.edu/S/primm .

C.3 Bayesian analysis

A considerable number of Bayesian software packages have emerged in the last ten to twenty years. The emergence rate steadily increases in tandem with advances in both computing speed and statistical methodology. Early attempts (from the 1970s and 80s) typically handled computer limitations by reducing the package's scope, often by limiting the dimension of the problem to be solved. Later, some packages began to focus on a particular class of models or problems, thus limiting the computational task somewhat while simultaneously heightening their appeal among practitioners in these areas. Some packages aim specifically at persons just learning Bayesian statistics (and perhaps computing as well), sacrificing breadth of coverage for clarity of purpose and ease of use. Most recently, a few general purpose Bayesian analysis engines have begun to emerge. Such generality essentially forces the package to employ MCMC computing methods, which as we know from Section 5.4 carry an array of implementational challenges.

A noteworthy early development in Bayesian computing was **sbayes**, a suite of *S* functions for approximate Bayesian analysis. Developed by Prof. Luke Tierney and recently updated by Prof. Alex Gitelman, **sbayes** computes approximate posterior moments and univariate marginal posterior densities via Laplace's method (second order approximation; see Subsection 5.2.2). First order (normal) approximation and importance sampling are also available as computational methods. The user must input an *S* function to compute the joint log-posterior (up to an additive constant), an initial guess for the posterior mode, optional scaling information, and the data. The program finds the true joint posterior mode via a modified

Newton algorithm, and then proceeds with Laplace's method. The program is freely available from *StatLib*, at

`http : //lib.stat.cmu.edu/S/sbayes/`

Alternatively, an *XLISP-STAT* version is available as part of the core *XLISP-STAT* system, an extendible environment for statistical computing and dynamic graphics based on the Lisp language. This system is also available from *StatLib* at

`http://lib.stat.cmu.edu/xlispstat/`

The Bayes functions in *XLISP-STAT* are documented in Section 2.8.3 of Tierney (1990).

Another recent program worth mentioning is **BAYESPACK,** Prof. Alan Genz's collection of Fortran routines for numerical integration and importance sampling. Based on methods described in Genz and Kass (1997), this collection is available from the website

`http://www.math.wsu.edu/math/faculty/genz/homepage`

C.3.1 Special purpose programs

Time series software

The **BATS** (*Bayesian Analysis of Time Series*) program focuses on the set of dynamic linear and nonlinear models, including Kalman filter models. Originally written in the APL language, it was developed at the University of Warwick and described in the paper by West, Harrison, and Pole (1987). Today, a substantially enhanced modular C version of **BATS** for Windows/DOS forms the computing basis for the popular Bayesian forecasting and time series book by Pole, West, and Harrison (1994). More information can found at

`http://www.isds.duke.edu/~mw/bats.html`

For those interested in non-stationary time series analysis, software and documentation for time-varying autoregressive modeling and time series decompositon analyses are available at

`http://www.isds.duke.edu/~mw/tvar.html`

This is a preliminary version of the software used for analyses in a series of recent articles by Drs. Raquel Prado and Mike West.

Relatedly, **ARCOMP** is a set of Fortran programs and *S-plus* functions that implement MCMC methods for autoregressive component modeling with priors on the characteristic roots. The approach involves focus on underlying latent component structure, and handles problems of uncertainty about model order and possible unit roots. This software by Drs. Gabriel Huerta and Mike West can be downloaded from

`http://www.isds.duke.edu/~mw/armodels.html`

Model selection software

Several *S-plus* functions for Bayesian model selection and averaging are freely available from *StatLib*, at

`http://lib.stat.cmu.edu/S/`

For instance, **bicreg**, **bic.logit** and **glib**, all written by Prof. Adrian Raftery of the University of Washington, use the BIC (Schwarz) criterion (2.21). The first does this in the context of linear regression models, the second for logistic regression models, and the third for generalized linear models; see Raftery (1996). Dr. Jennifer Hoeting offers **bma**, a program for model averaging in linear regression models. Finally, **bic.glm** and **bic.surv**, both contributed by Dr. Chris Volinsky, implement Bayesian model averaging using BIC, the first for generalized linear models and the second for Cox proportional hazards models. For more information, see also

`http://www.research.att.com/~volinsky/bma.html`

Software for clinical trials and other biomedical data

Spiegelhalter et al. (2000) list several websites concerned with methods and software for Bayesian methods in health technology assessment, especially clinical trial analysis. These include Prof. Lawrence Joseph's page,

`http://www.epi.mcgill.ca/~web2/Joseph/software.htm`

which gives a variety of programs for Bayesian sample size calculations, change-point methods, and diagnostic testing.

On the commercial side, there is **iBART** (*instructional Bayesian Analysis of Randomized Trials*), a new package for Windows 95, 98, and NT described on the web at

`http://www.talariainc.com/`

Loosely based on the original **BART** program written by Prof. David Spiegelhalter of the Medical Research Council (MRC) Biostatistics Unit at the University of Cambridge, **iBART** implements the Bayesian approach to clinical trial monitoring and analysis via an indifference zone, as described in Section 6.1.2 above. The software allows updating of normal and other more specialized survival likelihoods using standard prior distributions.

Finally, Prof. Giovanni Parmigiani has developed **BRCAPRO**, a Bayesian software program for the genetic counseling of women at high risk for hereditary breast and ovarian cancer. Available for PC and UNIX platforms, this program and related information is available at

`http://astor.som.jhmi.edu/brcapro/`

Belief analysis software

In the realm of belief analysis, Wooff (1992) describes a computer program called **[B/D]**, an acronym for *Beliefs adjusted by Data.* Developed at the Universities of Hull and Durham, the program implements the subjectivist Bayesian theory in which prevision (expectation), rather than probability, is the fundamental operation (see e.g. Goldstein, 1981). Given a partial specification of beliefs over a network of unknown quantities, the program checks them for internal consistency *(coherence)* and then adjusts each in the presence of the others and the observed data. **[B/D]** is written in Pascal, runs in either interactive or batch mode, and contains both progamming (e.g., if-then-else) and some statistical package (e.g., belief structure adjustment) elements. However, it is primarily a tool for organizing beliefs, rather than a statistical package for doing standard analyses (e.g., linear regression). The program is freely available at

`http://fourier.dur.ac.uk:8000/stats/bd/`.

A related program is the Bayesian Knowledge Discoverer, **BKD**, a computer program designed to extract Bayesian belief networks from (possibly incomplete) databases. Based on the "bound and collapse" algorithm, the aim of **BKD** is to provide a knowledge discovery tool able to extract reusable knowledge from databases. Also, the Robust Bayesian Classifier, **RoC**, is a program able to perform supervised Bayesian classification from incomplete databases, with no assumption about the pattern of missing data. Both programs are available at the Open University (U.K.) website

`http://kmi.open.ac.uk/projects/bkd/`

Expert systems software

An important recent application of Bayesian methodology is to the field of *expert systems,* since automatic updating of the information available to a decision-maker given the current collection of data is most simply and naturally done via Bayes' Rule. Before describing the software packages available for this type of analysis, we must first briefly describe the modeling framework in which they operate.

The probability and conditional independence structure of the variables within an expert system is often represented pictorially in a *graphical model.* In such a graph, *nodes* in the graph represent variables while *directed edges* between two nodes represent direct influence of one variable on the other. As a result, missing links represent irrelevance between nodes. To avoid inconsistencies, we do not allow a sequence of directed edges in the graph that returns to its starting node (a *cycle*), so our graph is a *directed acyclic graph*, or *DAG*. Note that such a graph may be constructed in a given problem setting without reference to probability, since it requires only intuition about the causal structure in the problem. As such, DAGs are often easily specified by practitioners, and helpful in organizing thinking about the problem before getting bogged down in quantitative details. Spiegelhalter et al. (1993) provide a thorough review of developments in the use of graphical models and Bayesian updating in expert systems.

To add the probability structure to the graphical model, we write $V = \{v\}$ as the collection of nodes in the graph. Then redefining "irrelevance" as "conditional independence," we have that

$$p(V) = \prod_{v \in V} p(v|\text{pa}(v)) \ , \tag{C.1}$$

where pa(v) denotes the *parent* nodes (direct antecedents) of node v. Thus, the probability structure of the graph is determined by "prior" distributions for the nodes having no parents, and the conditional distributions for the remaining "children" given their parents. The marginal distribution of any node given any observed "data" nodes could then be found via Bayes' Rule.

In the case where all of the v are given *discrete* distributions, the Bayesian updating can be carried out exactly using the algorithm of Lauritzen and Spiegelhalter (1988). This algorithm is perhaps best described as a "two-pass" algorithm, as the node tree is traversed in both directions when "collecting' and "distributing" evidence. Perhaps the most professional piece of software implementing this algorithm is **Hugin**, a Danish commercial product distributed by Hugin Expert A/S. (Incidentally and somewhat refreshingly, **Hugin** is not an acronym, but rather the name of one of two ravens in Scandinavian mythology who brought back news from every corner of the world to the god Odin.) The newest versions of the program allow *influence diagrams* (a graphical probability model augmented with decision and utilities) and *chain graphs* (a graphical model with undirected, as well as directed. edges); future enhancements will include parametric learning via the EM algorithm. Further information on obtaining **Hugin** can be obtained from the company's homepage,

http://hugin.dk/

Freeware for evaluation and updating of expert systems models is also available from several sources. A remarkably comprehensive list of the commercial and noncommercial software for processing probabilistic graphical models (as well as more general kinds of graphical belief function models) is provided by Dr. Russell Almond at the webpage

http://bayes.stat.Washington.edu/almond/belief.html/

Finally, the members of the Decision Theory and Adaptive Systems

Group at Microsoft Research have developed a Bayesian belief network construction and evaluation tool called **MSBN**. It supports the creation and manipulation of Bayesian networks as well as their evaluation. Complete information about **MSBN** and how to obtain it can be downloaded from the website

http://fwww.research.microsoft.com/dtas/msbn/default.htm

The program and its component parts are available free of charge for academic and commercial research and education following the completion of an on-line licensing agreement.

Neural networks software

Neural networks have a close relation to expert systems. Algorithms for Bayesian learning in these models using MCMC methods were described by Neal (1996b). Software implementing these algorithms is freely available over the web at

http://www.cs.toronto.edu/~radford/fbm.sofware.html

The software supports models for regression and classification problems based on neural networks and Gaussian processes, and Bayesian mixture models. It also supports a variety of Markov chain sampling methods.

C.3.2 Teaching programs

First Bayes

Written by Prof. Tony O'Hagan, **First Bayes** is a user-friendly package to help students to learn and understand elementary Bayesian statistics. Particular features include

- Plotting and summarization of thirteen standard distributional families, and mixtures of these families;
- Analysis of the one-way ANOVA and simple linear regression models, with marginal distributions for arbitrary linear combinations of the location parameters;
- Automatic updating of posterior and predictive summaries following user-made changes to the data or prior distribution;
- Sequential analysis, via a facility for setting the current posterior equal to the prior.

The program is written in APL, a language to which the author displays the extreme devotion typical of APL users. It runs under Windows 3.1, 95, and NT, and may be freely used and copied for any noncommercial purposes. It may be obtained over the web directly from the author at

http://www.shef.ac.uk/~st1ao/1b.html/

Minicab macros

A recent book by Albert (1996) describes a set of macros written in Minitab, an easy-to-use statistical package widely used in U.S. elementary statistics courses. The macros are designed to perform all of the computations in Berry (1996), an elementary statistics text that emphasizes Bayesian methods. Included are programs to

- Simulate games of chance;
- Perform inference for one- and two-sample proportion (binomial) problems;
- Perform inference for one- and two-sample mean (normal) problems;
- Perform inference for the simple linear regression model;
- Illustrate Bayes tests and predictive inference;
- Summarize arbitrary one- and two-parameter posterior distributions using a variety of integration strategies, including sampling-importance resampling (weighted bootstrap), Laplace's method, adaptive quadrature, and the Gibbs and Metropolis samplers.

Information is available with the two textbooks mentioned above, or directly from the author via the website

http://bayes.bgsu.edu/~albert/mini_bayes/info.html .

C.3.3 Markov chain Monte Carlo programs

Tools for convergence diagnosis

Before describing the packages available for MCMC sampling-based Bayesian analysis, we provide information on obtaining the programs mentioned near the end of Chapter 5 for assessing whether the output from an MCMC sampler suggests convergence of the underlying process to stationarity. These programs generate no samples of their own, but merely operate on the output of a user-provided MCMC algorithm.

First, the *S* program created by Andrew Gelman to implement the scale reduction factor (5.46) advocated as a convergence diagnostic by Gelman and Rubin (1992) is available over the web from *StatLib* at

http://lib.stat.cmu.edu/S/itsim .

The code computes both the mean and upper 95^{th} percentile of the scale reduction factor, which must approach 1 as $n \rightarrow \infty$.

Second, Steven Lewis has written an implementation of the Raftery and Lewis (1992) diagnostic procedure, which attempts to find the number of samples from a single MCMC chain needed to estimate some posterior quantile to within some accuracy with some high probability. A Fortran version of this code is available over the web from *StatLib* at

`http://lib.stat.cmu.edu/general/gibbsit` ,

while an *S* translation is located in

`http://lib.stat.cmu.edu/S/gibbsit` .

Both of the above *S* programs, as well as several other convergence diagnostics, are also implemented in the **CODA**, **BOA**, and **WinBUGS** packages; see below.

BUGS, WinBUGS, CODA, and BOA

The easy accessibility and broad applicability of the sampling based approach to Bayesian inference suggests the feasibility of a general-purpose language for Bayesian analysis. Until recently, most of the MCMC computer code written was application-specific: the user simply wrote a short piece of Gibbs or Metropolis sampling code for the problem at hand, and modified it to fit whatever subsequent problems came along. Practitioners have since recognized that certain high-level languages (such as *S* and *XLISP-STAT)* are quite convenient for the data entry. graphical convergence monitoring, and posterior summary aspects of the analysis, while lower-level, compiled languages (such as Fortran or C) are often needed to facilitate the enormous amount of random generation and associated looping in their sampling algorithms. Of the two tasks. the lower-level portion is clearly the more difficult one, since the computer must be taught how to understand a statistical model, organize the prior and likelihood components, and build the necessary sampling distributions before sampling can actually begin.

In this vein, the most general language developed thus far seems to be **BUGS**, a clever near-acronym for *Bayesian inference Using Gibbs Sampling.* Developed at the MRC Biostatistics Unit at the University of Cambridge and initially described by Gilks, Thomas, and Spiegelhalter (1992), **BUGS** uses *S*-like syntax for specifying hierarchical models. The program then determines the complete conditional distributions necessary for the Gibbs algorithm by converting this syntax into a directed acyclic graph (see Sub-

eters in the model. **BUGS** successively samples from the parameter nodes, writing the output to a file for subsequent convergence assessment and posterior summarization.

To see how **BUGS** determines the appropriate sampling chain, note that the full conditional distribution for each node can be inferred from the links in the graph and the conditional distribution at each node. That is, in the notation of equation (C.1),

$$p(v \mid V \backslash v) \propto p(v, V \backslash v) = p(V) ,$$

where $V \backslash v$ (read "V remove v") is the collection of all variables except for v. This distribution is proportional to the product of the terms in $p(V)$ containing v, i.e.,

$$p(v|V \backslash v) \ \propto \ p(v|\mathrm{pa}(v)) \prod_{v \in \mathrm{pa}(w)} p(w|\mathrm{pa}(w)) \ .$$

The first term on the right-hand side is the distribution of v given its parents, while the second is the product of distributions in which v is itself a parent. Thus, cycling through the node tree in the usual Gibbs updating fashion (see Subsection 5.4.2) produces a sample from the full joint distribution of V at convergence.

Recall that if all of the variables v can assume only a finite number of values, then exact, noniterative algorithms based on the work of Lauritzen and Spiegelhalter (1988) are available for computing the posterior distribution of any node in the graph. If some of the v are continuous, however, no such algorithm is available, but the preceding discussion shows the problem is still natural for Gibbs sampling. **BUGS** thus affords much more general graphical models than, say, **Hugin**, but also has the disadvantage of being both iterative and approximate, due to its MCMC nature.

BUGS is able to recognize a number of conjugate specifications (e.g., normal/normal, Poisson/gamma, etc.), so that the corresponding nodes can be updated quickly via direct sampling. Adaptive rejection sampling (Gilks and Wild, 1992) is used for other nodes having log-concave full conditionals; the remaining nodes can be updated via Metropolis sampling, or else discretized and their standardized full conditionals computed by brute force (as would be done for any discrete-finite node). Other helpful features include a variety of distributions (including the multivariate normal and Wishart), functions (including the log-gamma and matrix inverse), input capabilities (to facilitate multiple sampling chains or multiple priors), and advice on implementing a variety of Bayesian data analysis tools, such as reparametrization, model criticism, and model selection. An extensive examples manual is also available.

Recently, development of the **BUGS** language has focused on its Windows version, **WinBUGS**. While "classic" **BUGS** uses only text representations of models and commands to drive the simulation, **WinBUGS** can also use a graphical interface called **DoodleBUGS** to construct the model, and a point-and-click Windows interface for controlling the analysis. There are also graphical tools for monitoring convergence of the simulation. The model specification and simulation are separate components: text versions of the model can be output from **DoodleBUGS**, and text model descriptions can also be used as input to the simulation program. **WinBUGS** also includes several features not part of classic **BUGS**, including a more general Metropolis sampler, on-line graphical monitoring tools, space-efficient monitoring of running means and standard deviations, some automatic residual plots for repeated measures data, and capability for multiple sampling chains and Brooks-Gelman-Rubin diagnostics (see Brooks and Gelman, 1998). All the documentation specific to **WinBUGS** is packaged with the program.

The most recent extensions to **WinBUGS** are **PKBUGS**, a version specifically tuned for pharmacokinetic (complex nonlinear) modeling, and **GeoBUGS**, which allows the the fitting and displaying of spatial models. **GeoBUGS** will ultimately include an importer for maps in *Arc View*, *S-plus*, and *EpiMap* format.

A more extensive "back end" for convergence diagnosis or posterior summarization is provided by **CODA** (a musical analogy, and also an anagram of *Convergence Diagnosis and Output Analysis*). **CODA** is a menu-driven suite of *S-plus* functions for investigating MCMC output in general, and **BUGS** output in particular. The program enables

- Various summary statistics, such as means, quantiles, naive and batched standard errors, and Geweke (1992) "numerical standard errors;"
- sample traces and density plots;
- sample autocorrelations (within a chain for one parameter) and cross-correlations (across one or more chains for two parameters);
- Several other popular statistics and plots for convergence diagnosis (see Subsection 5.4.6), including those of Gelman and Rubin (1992), Geweke (1992), Raftery and Lewis (1992), and Heidelberger and Welch (1983).

Originally written by Prof. Kate Cowles while still a graduate student in the Division of Biostatistics at the University of Minnesota, **CODA** source code is now maintained and distributed along with **BUGS** and **WinBUGS**. **CODA** allows easy analysis of multiple sampling chains, changing of default settings for plots and analysis, and retaining of text and graphical output for subsequent display (indeed, several of the plots in this book were created in **CODA**) . The program runs on the same platforms as **BUGS** and **WinBUGS**, and source code is available.

BUGS and **CODA** each come with comprehensive, easy-to-read user guides (Spiegelhalter et al., 1995a, and Best, Cowles, and Vines, 1995, respectively). A booklet of worked **BUGS** examples is also available (Spiegelhalter et al., 1995b). The programs and manuals are available over the web at

http : //www.mrc-bsu.cam. ac.uk/bugs/ .

In the future, the programs will be shareware, meaning that a nominal fee will be requested of those users requiring support.

Finally, Brian Smith of the University of Iowa has written an *S-plus*/*R* revision of **CODA** called **BOA** (*Bayesian Output Analysis*) for carrying out convergence diagnostics and statistical and graphical analysis of Monte

Carlo sampling output. **BOA** can be used as an output processor for **BUGS**, **WinBUGS**, or any other program which produces sampling output. It is quite stable, easy to use, feature rich, and freely available over the web at

http://www.public-health.uiowa.edu/boa/

bpois

Doss and Narasimhan (1994) describe a Gibbs sampling program for Bayesian Poisson regression called **bpois**. The program combines Gibbs sampling with a graphical interface in the *LISP-STAT* programming language. It also enables sensitivity analysis via importance sampling with dynamic graphics. The program is fully menu-driven, and features dialog boxes with sliding control mechanisms to change the relevant hyperparameter values. The code is specific to Poisson likelihoods, but could be customized by the user willing to write the necessary *C* subroutines. The program is freely available from *StatLib* via the web at

http://lib.stat.cmu.edu/xlispstat/bpois

Of course, the user must already have access to the core *LISP-STAT* system; see the instructions above.

MCSim

Originally developed to do uncertainty analysis (simple Monte Carlo simulations) on models specified by ordinary differential equations (e.g., pharmacokinetic models), **MCSim** is a simulation package that allows the user to design statistical or simulation models, run Monte Carlo stochastic simulations, and perform Bayesian inference through MCMC simulations. The program is written in *C*, with both source code and a user's manual available. The program and further information is available from the webpage maintained by the program's primary author, Frederic Bois, at

http://sparky.berkeley.edu/users/fbois/mcsim.html

BINGO

An acronym for *Bayesian Inference - Numerical, Graphical and Other stuff*, this program is an extended version of **Bayes Four**, developed by Prof. Ewart Shaw at Warwick University. Built for portable APL workspaces, an initial version is available at

http://www.Warwick.ac.uk/statsdept/Staff/JEHS/index.htm

BINGO features a "multi-kernel" version of the Naylor and Smith (1982) algorithm, highly efficient ("designed") numerical integration rules, device-independent graphics, and facilities for other approaches. The user must supply APL code for the likelihood and so on, similar to **Bayes Four** Other facilities include built-in reparametrizations, Monte Carlo importance sampling and the Metropolis-Hastings algorithm, quasirandom sequences, Laplace approximations, function maximization, APLTEX, and the compact device-independent graphics language WGL (*Warwick Graphics Language,* pronounced "wiggle").

APPENDIX D

Answers to selected exercises

Chapter 1

1. Though we may have no prior understanding of male hypertension, the structure of the problem dictates that $\theta \in [0, 1]$, providing some information. In addition, values close to each end of this interval seem less likely, since we know that while not *all* males are hypertensive, at least *some* of them are.

   Suppose we start with $\hat{\theta} = .3$. Given the knowledge that 4 of the first 5 men we see are in fact hypertensive, we would likely either (a) keep our estimate unchanged, since too few data have yet accumulated to sway us, or (b) revise our initial estimate upward somewhat - say, to .35. (It certainly would not seem plausible to revise our estimate *downward* at this point.)

   Once we have observed 400 hypertensive men out of a sample of 1000, however, the data's pull on our thinking is hard to resist. Indeed, we might completely discard our poorly-informed initial estimate in favor of the data-based estimate, $\hat{\theta} = .4$.

   All of the above characteristics of sensible data analysis (plausible starting estimates, gradual revision, and eventual convergence to the data-based value) are features we would want our inferential system to possess. These features *are* possessed by Bayes and empirical Bayes systems.

5.(a) p-value = $P(X \geq 13 \text{ or } X \leq 4) = .049$ under H_0. So, using the usual Type I error rate $\alpha = .05$, we would stop and reject H_0.

(b) For the two-stage design, we now have

$$\begin{aligned} p-\text{value} &= P(X_1 \geq 13 \text{ or } X_1 \leq 4) \\ &\quad +P(X_1 + X_2 \geq 29 \text{ and } 4 < X_1 < 13) \\ &\quad +P(X_1 + X_2 \leq 15 \text{ and } 4 < X_1 < 13) \\ &= .085\ , \end{aligned}$$

which is no longer significant at $\alpha = .05$! Thus even though the observed data was exactly the same, the mere *contemplation* of a second stage to the experiment (having no effect on the data) changes the answer in the classical hypothesis testing framework.

(c) The impact of adding imaginary stages to the design could be that the p-value approaches 1 even as x_1 remains fixed at 13!

(d) Stricly speaking, since this aspect of the design (early stopping due to ill effects on the patients) wasn't anticipated, the p-value is not computable!

(e) To us it suggests that p-values are less than objective evidence, since not only the form of the likelihood, but also extra information (like the design of the experiment), is critical to their computation and interpretation.

6.(a) $B = .5$, so $\theta|y \sim N(.5\mu + .5y, .5(2)) = N(2,1)$.

(b) Now $B = .1$, so $\theta|y \sim N(.1\mu + .9y, .9(2)) = N(3.6, 1.8)$. The increased uncertainty in the prior produces a posterior that is much more strongly dominated by the likelihood; the posterior mean shrinks only 10% of the way toward the prior mean μ. The overall precision in the posterior is also much lower, reflecting our higher overall uncertainty regarding θ. Despite this decreased precision, a frequentist statistician might well prefer this prior, since it allows the data to drive the analysis. It is also likely to be more robust, i.e., to perform well over a broader class of θ-estimation settings.

Chapter 2

3. From Appendix Section A.2, since $\chi^2(\nu) \equiv G(\nu/2, 2)$, we have that $G(\alpha, 2) \equiv \chi^2(2\alpha)$. But if $X \sim G(\alpha, 2)$, a simple Jacobian transformation shows that $Y = \beta X/2 \sim G(\alpha, \beta)$. Finally, $Z = 1/Y \sim IG(\alpha, \beta)$ by definition, so we have that the necessary samples can be generated as

$$\sigma^2 = \frac{2}{\beta X}, \text{ where } X \sim \chi^2(2\alpha)$$

and

$$\alpha = \frac{1}{2} + a \ , \quad \beta = \left[\frac{1}{2}(x-\theta)^2 + \frac{1}{b}\right]^{-1}$$

4. Since $\Theta = \Re^+$ in this example, a difficulty with the histogram approach is its finite support (though in this example, terminating the histogram at, say, ten feet would certainly seem safe). Matching a parametric form avoids this problem and also streamlines the specification process, but an appropriate shape for a population maximum may not be immediately obvious, nor may the usual families be broad enough to truly capture your prior beliefs.

5.(a) $\theta \sim Beta(\alpha, \beta)$

(b) $\theta \sim Beta(\alpha, \beta)$ (likelihood same as in (a))

(c) $\boldsymbol{\theta} \sim D(\boldsymbol{\alpha})$

(d) $\beta \sim IG(a, b)$

(e) none available

(f) $\boldsymbol{\theta} \sim N_k(\boldsymbol{\mu}, \mathbf{V})$

(g) $\boldsymbol{\Sigma} \sim W(\boldsymbol{\Omega}, \nu)$

6.(a) scale; $p(\sigma) \propto 1/\sigma$

(b) location-scale; $p(\theta, a) \propto 1/a$

(c) location; $p(\mu) \propto 1$

(d) scale; $p(\beta) \propto 1/\beta$

(e) neither (α is often called a *shape* parameter)

7. The Jeffreys prior for γ is

$$I(\gamma) = -E_{\mathbf{x}|\gamma}\left[\frac{d^2}{d\gamma^2}\log p(\mathbf{x}|\gamma)\right] .$$

But, using the chain rule, the product rule, and the chain rule again, we can show that

$$\frac{d^2}{d\gamma^2}\log p(\mathbf{x}|\gamma) = \frac{d\log p(\mathbf{x}|\theta)}{d\theta}\cdot\frac{d^2\theta}{d\gamma^2} + \frac{d^2\log p(\mathbf{x}|\theta)}{d\theta^2}\cdot\left(\frac{d\theta}{d\gamma}\right)^2 .$$

Since the second terms in these two products are constant with respect to x and the score statistic has expectation 0, we have

$$\begin{aligned} I(\gamma) &= -E_{\mathbf{x}|\theta}\left[\frac{d^2\log p(\mathbf{x}|\theta)}{d\theta^2}\right]\cdot\left(\frac{d\theta}{d\gamma}\right)^2 \\ &= I(\theta)\cdot\left(\frac{d\theta}{d\gamma}\right)^2 . \end{aligned}$$

Hence $[I(\gamma)]^{1/2} = [I(\theta)]^{1/2}\cdot|d\theta/d\gamma|$, as required.

10.(a) For the joint posterior of θ and σ^2, we have

$$p(\theta, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-n/2-1}\exp\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\} . \qquad \text{(D.1)}$$

This expression is proportional to an inverse gamma density, so integrating it with respect to σ^2 produces

$$p(\theta|\mathbf{y}) \propto \left[\sum_{i=1}^{n}(y_i-\theta)^2\right]^{-n/2}$$

But using the identity $\sum_{i=1}^{n}(y_i-\theta)^2 = C + n(\bar{y}-\theta)^2$, where $C = \sum_{i=1}^{n}(y_i-\bar{y})^2$, we have that

$$p(\theta|\mathbf{y}) \propto [C + n(\bar{y}-\theta)^2]^{-n/2} \propto [1 + t^2/(n-1)]^{-n/2} .$$

(b) Here, the above identity must be employed right away, in order to integrate θ out of the joint posterior (D.1) as proportional to a normal form. The result for $p(\sigma^2|\mathbf{y})$ is then virtually immediate.

Chapter 3

1.(a) For $f(y|\theta) = N(y|\theta, 1)$, we have that $\theta = y + f'(x|\theta)/f(x|\theta)$. The result follows from substituting this expression for θ_i in the numerator of the basic posterior mean formula,

$$E(\theta_i|y_i) = \frac{\int \theta_i f(x_i|\theta_i) dG(\theta_i)}{\int f(x_i|\theta_i) dG(\theta_i)} ,$$

and simplifying.

(*Note:* The solution to a more general version of this problem is given by Maritz and Lwin, 1989, p.73.)

2. The computational significance of this result is that we may find the joint marginal density of y (formerly a k-dimensional integral with respect to $\boldsymbol{\theta}$ as the product over i of $m(y_i)$, the result of a one-dimensional integral. Since $m(\mathbf{y}|\eta)$ must emerge in closed form for computation of a marginal MLE $\hat{\eta}$ to be feasible, this reduction in dimensionality plays an important role in PEB analysis.

3.(a) $m(y_i|\eta) = \Gamma(\alpha+\eta)y_i^{\alpha-1}/[\Gamma(\alpha)\Gamma(\eta)(y_i+1)^{\alpha+\eta}]$

(b) With $\alpha = 2$,

$$L(\eta|\mathbf{y}) \propto \prod_{i=1}^{k} \frac{\eta(\eta+1)y_i}{(y_i+1)^{\eta+2}} ,$$

Taking the derivative of the log-likelihood with respect to η and setting the result equal to 0, we have that the marginal MLE $\hat{\eta}$ is a positive root of the quadratic equation

$$S\eta^2 + (S - 2k)\eta - k = 0 ,$$

where $S = \sum_{i=1}^{k} \log(y_i + 1)$.

4.(a) The marginal loglikelihood for B is proportional to

$$\frac{k}{2} \log B - \frac{B}{2} \sum_{i=1}^{k} y_i^2 .$$

Taking the derivative with respect to B, setting equal to 0, solving, and remembering the restriction that $0 \leq B \leq 1$, we have that

$$\widehat{B}_{MLE} = \min\left(\frac{k}{\sum_{i=1}^{k} y_i^2} , 1 \right) .$$

The resulting PEB point estimates are $\hat{\theta}_i^{EB} = (1 - \widehat{B}_{MLE})Y_i$.

(b) Since $\sum_{i=1}^{k}(\sqrt{B}Y_i)^2 = B||\mathbf{Y}||^2 \sim \chi^2(k) = G(k/2,2)$, it follows that $||\mathbf{Y}||^2 \sim G(k/2, 2/B)$, and hence that $1/(||\mathbf{Y}||^2) \sim IG(k/2, 2/B)$. Thus we can compute

$$E(\widehat{B}) = E\left(\frac{k-2}{||\mathbf{Y}||^2}\right) = \frac{k-2}{\frac{2}{B}\left(\frac{k}{2}-1\right)} = \frac{B(k-2)}{k-2} = B$$

and so $\hat{B}$ is indeed unbiased for B. But then

$$\hat{\theta}_i^{EB} = (1-\widehat{B})Y_i = \left(1 - \frac{k-2}{||\mathbf{Y}||^2}\right)Y_i = \hat{\theta}_i^{JS}$$

6. We need to compute

$$E_{\mathbf{Y}|\boldsymbol{\theta}}\left[\frac{1}{||\ \mathbf{Y}\ ||^2}\right].$$

Consider the case where $\sigma^2 = 1$ and replace $\mathbf{Y}$ by $\mathbf{Z}$; the general case follows by writing $\mathbf{Y} = \sigma\mathbf{Z}$. Now $||\ \mathbf{Z}\ ||^2$ is distributed as a non-central chi-square on k df with non-centrality parameter $\gamma = ||\ \boldsymbol{\theta}\ ||^2/2$. But a non-central chi-square can be represented as a discrete mixture of central chi-squares, i.e., $\chi^2(k+2J)$ where $J \sim Po(\gamma)$. Therefore,

$$\begin{aligned} E_{\mathbf{Z}|\boldsymbol{\theta}}\left[\frac{1}{||\ \mathbf{Z}\ ||^2}\right] &= E_{\mathbf{Z}|\gamma}\left[\frac{1}{||\ \mathbf{Z}\ ||^2}\right] \\ &= \sum_{j=0}^{\infty}\frac{\gamma^j}{j!}e^{-\gamma}\frac{1}{k+2j-2}, \end{aligned}$$

since $1/\chi^2(k+2J) \equiv IG((k+2j)/2, 2)$ (see Appendix Section A.2).

9. The population moments in the marginal family $m(r_i|\alpha,\beta)$ are

$$E(r_i) = \frac{1}{t_i}E(Y_i) = \frac{1}{t_i}E(E(Y_i|\theta_i)) = \frac{1}{t_i}E(t_i\theta_i) = E(\theta_i) = \alpha\beta$$

and

$$\begin{aligned} Var(r_i) &= \tfrac{1}{t_i^2}Var(Y_i) \\ &= \tfrac{1}{t_i^2}[Var(E(Y_i|\theta_i)) + E(Var(Y_i|\theta_i))] \\ &= \tfrac{1}{t_i^2}[Var(t_i\theta_i) + E(t_i\theta_i)] \\ &= \tfrac{1}{t_i^2}[t_i^2(\alpha\beta^2) + t_i(\alpha\beta)] = \alpha\beta^2 + t_i^{-1}\alpha\beta\ . \end{aligned}$$

This results in the system of 2 equations and 2 unknowns

$$\begin{gathered} \bar{r} = \alpha\beta \\ s_r^2 = \alpha\beta^2 + \alpha\beta\textstyle\sum_{i=1}^{k} t_i^{-1}/k \end{gathered}.$$

Solving, we obtain $\hat{\alpha} = \bar{r}^2/(s_r^2 - \bar{r}\sum_{i=1}^{k} t_i^{-1}/k)$ and $\hat{\beta} = \bar{r}/\hat{\alpha}$.

Chapter 4

2.(a) The MLE is equal to the Bayes rule when $M = 0$, and carries a risk (MSE) of $E_{X|\theta}(\frac{X}{n} - \theta)^2 = Var_{X|\theta}(\frac{X}{n}) = \theta(1-\theta)/n$. Subtracting the risk of the Bayes estimate given in equation (4.4) and setting $\mu = \frac{1}{2}$, we have risk improvement by the Bayes estimate if and only if

$$\frac{\theta(1-\theta)}{n}\left[1 - \left(\frac{n}{M+n}\right)^2\right] - \left(\frac{M}{M+n}\right)^2\left(\theta - \frac{1}{2}\right)^2 \geq 0 . \quad \text{(D.2)}$$

The two θ values for which this expression equals 0 (available from the quadratic formula) are the lower and upper bounds of the risk improvement interval. Clearly this interval will be symmetric about $\theta = \frac{1}{2}$, since there the first term in (D.2) is maximized while the second is minimized (equal to 0).

(b) Since $\theta|x \sim Beta(a+x, b+n-x)$, we have

$$\begin{aligned}
\hat{\theta}_{Bayes} &= \frac{\int \frac{\theta}{\theta(1-\theta)}\,\theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta}{\int \frac{1}{\theta(1-\theta)}\,\theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta} \\
&= \frac{\int \theta^{a+x-1}(1-\theta)^{b+n-x-2}d\theta}{\int \theta^{a+x-2}(1-\theta)^{b+n-x-2}d\theta} \\
&= \frac{\frac{\Gamma(a+x)\Gamma(b+n-x-1)}{\Gamma(a+b+n-1)}}{\frac{\Gamma(a+x-1)\Gamma(b+n-x-1)}{\Gamma(a+b+n-2)}} \\
&= \frac{a+x-1}{a+b+n-1}
\end{aligned}$$

from the result of Appendix B, Problem 1.

6.(b) This is actually the prior used by Müller and Parmigiani (1995). Using the "shortcut" method (sampling both θ_j^* and y_j^* with $N = 1$), they obtain $\tilde{n} = 29$.

Even this mixture prior is amenable to a fully analytical solution, since the posterior distribution is still available as a mixture of beta densities. The authors then obtain the refined estimate $\tilde{n} = 34$.

Chapter 5

3. The probability p of accepting a rejection candidate is

$$\begin{aligned}
p &= P\left(L(\theta)\pi(\theta) > MUg(\theta)\right) \\
&= \int P\left(U < \frac{L(\theta)\pi(\theta)}{Mg(\theta)}\right) g(\theta)d\theta \\
&= \int \frac{L(\theta)\pi(\theta)}{Mg(\theta)} \cdot g(\theta)d\theta
\end{aligned}$$

$$= \frac{1}{M}\int L(\theta)\pi(\theta)d\theta$$
$$= \frac{c}{M}\,.$$

4. The new posterior is

$$p_2(\theta|\mathbf{y}) \propto \frac{\pi_2(\theta)}{\pi_1(\theta)}p_1(\theta|\mathbf{y})\,.$$

In the notation of the weighted bootstrap, we have $g(\theta) = p_1(\theta|\mathbf{y})$, and we seek samples from

$$L(\theta)\pi(\theta) = \frac{\pi_2(\theta)}{\pi_1(\theta)}p_1(\theta|\mathbf{y}) \quad \Rightarrow \quad w(\theta) = \frac{\pi_2(\theta)}{\pi_1(\theta)}$$

Thus if we compute

$$q_i = \frac{\pi_2(\theta_i)/\pi_1(\theta_i)}{\sum_{j=1}^N \pi_2(\theta_j)/\pi_1(\theta_j)}$$

then resampling θ_i^*s with probability qi provides a sample from $p_2(\theta|\mathbf{y})$. This procedure is likely to work well only if π_2 has lighter tails than π_1.

15. The Hastings acceptance ratio *r* is

$$\begin{aligned} r &= \frac{p(v)q(u,\, v)}{p(u)q(v,\, u)} \\ &= \frac{p(v)p(u_i|u_{j\neq i})}{p(u)p(v_i|u_{j\neq i})} \\ &= \frac{p(v_i|u_{j\neq i})p(u_{j\neq i})p(u_i|u_{j\neq i})}{p(u_i|u_{j\neq i})p(u_{j\neq i})p(v_i|u_{j\neq i})} \\ &\equiv 1\,, \end{aligned}$$

since *v* and *u* differ only in their i^{th} components.

17.(a) Since $\log p_Z(z) \propto (\delta-1)\log z - \beta z + k\log(1-e^{-z})$, we have

$$\frac{\partial \log p_Z(z)}{\partial z} \propto \frac{\delta-1}{z} - \beta + \frac{k}{e^z-1}$$

and

$$\frac{\partial^2 \log p_Z(z)}{\partial z^2} \propto \frac{-(\delta-1)}{z^2} - \frac{ke^z}{(e^z-1)^2} < 0 \text{ for } \delta \geq 1\,.$$

Thus $p_Z(z)$ is guaranteed to be log-concave provided $\delta \geq 1$.

(b) Marginalizing the joint density, we have

$$\int p_{Z,U}(z,u)du \quad \propto \quad \int \cdots \int z^{\delta-1}e^{-\beta z}\prod_{i=1}^k I_{(e^{-z},1)}(u_i)du_1\cdots du_k$$

$$
\begin{aligned}
&= z^{\delta-1}e^{-\beta z}\prod_{i=1}^{k}\int I_{(e^{-z},1)}(u_i)du_i \\
&= z^{\delta-1}e^{-\beta z}\prod_{i=1}^{k}(1-e^{-z}) \\
&= z^{\delta-1}e^{-\beta z}(1-e^{-z})^k \;\propto\; p_Z(z)
\end{aligned}
$$

(c) The full conditional for U is

$$p_{U|Z}(u|z) \propto p_{U,Z}(u,z) \propto \prod_{i=1}^{k} I_{(e^{-z},1)}(u_i) = \prod_{i=1}^{k} Unif(e^{-z},1) \ ,$$

i.e., conditionally independent uniform distributions. For Z, we have

$$p_{Z|U}(z|u) \propto z^{\delta-1}e^{-\beta z} I_{(e^{-z},1)}(u_1)\cdots I_{(e^{-z},1)}(u_k) \ .$$

So $e^{-z} < u_i$ for all i implies $z > -\log u_i$ for all i, which in turn implies $z > -\log u_{min}$ where $u_{min} = \min(u_1,\ldots,u_k)$. Hence,

$$p_{Z|U}(z|u) \propto z^{\delta-1}e^{-\beta z} I_{(-\log u_{min}\,,\,\infty)}(z) \ ,$$

a *truncated gamma* distribution. Alternately sampling from $p_{U|Z}$ and $p_{Z|U}$ produces a Gibbs sampler which converges to $p_{U,Z}$; peeling off the $Z^{(g)}$ draws after convergence yields the desired sample from the D-distribution p_Z. Special subroutines required to implement this algorithm would include a $Unif(0,1)$ generator, so that the u_i draws could be computed as $e^{-z} + (1-e^{-z}) \cdot Unif(0,1)$, and a gamma cdf and inverse cdf routine, so that the z draws could be found using the truncated distribution formula given in Chapter 7, problem 4.

19. We have that

$$
\begin{aligned}
A_{x|x} &= A_{y|x}A_{x|y} \\
&= \begin{pmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{pmatrix} \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix} \\
&= \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix} .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
f_x A_{x|x} &= \begin{pmatrix} p_1+p_2 & p_3+p_4 \end{pmatrix} \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix} \\
&= \begin{pmatrix} p_1+p_3 & p_2+p_4 \end{pmatrix} \\
&= f_x \ .
\end{aligned}
$$

20.(b) Analytically, we have

$$\begin{aligned} f_x(x) &= \int_0^\infty \left[\int_0^\infty y e^{-yx} t e^{-ty} dy\right] f_x(t)dt \\ &= \int_0^\infty \left[\frac{t}{(x+t)^2}\right] f_x(t)dt \end{aligned}$$

$f_x(\text{t}) = 1/t$ is a solution to this equation. But no Gibbs convergence is possible here, since this is not a density function; $\int_0^\infty f_x(t)dt = +\infty$. The trouble is that the complete conditionals must determine a *proper* joint density.

22.(a) $p(\theta_1|\theta_2, y)$ and $p(\theta_2|\theta_2, y)$ are

$$N\left(\frac{b_1^2(y-\theta_2)+a_1}{1+b_1^2}\,,\ \frac{b_1^2}{1+b_1^2}\right)$$

and

$$N\left(\frac{b_2^2(y-\theta_1)+a_2}{1+b_2^2}\,,\ \frac{b_2^2}{1+b_2^2}\right)\ ,$$

respectively.

(b) $p(\theta_1|y)$ and $p(\theta_2|y)$ are

$$N\left(\frac{b_1^2(y-a_2)+(1+b_2^2)a_1}{1+b_1^2+b_2^2}\,,\ \frac{b_1^2(1+b_2^2)}{1+b_1^2+b_2^2}\right)$$

and

$$N\left(\frac{b_2^2(y-a_1)+(1+b_1^2)a_2}{1+b_1^2+b_2^2}\,,\ \frac{b_2^2(1+b_1^2)}{1+b_1^2+b_2^2}\right)\ ,$$

respectively, so the data do indeed inform (i.e., there is "posterior learning").

(c) Answer provided in Figure D.1. Convergence is obtained immediately for μ, but never for θ_1 or θ_2. Note that the estimates of $E(\mu|y)$ at 100 and 1000 iterations are numerically identical.

(d) Answer provided in Figure D.2. Now we see very slow convergence for μ, due to the very slow convergence for the (barely identified) θ_1 and θ_2. The estimates of $E(\mu|y)$ at 100 and 1000 iterations are now appreciably different (with only the latter being correct), but there is no indication in the 100 iteration μ plot, G&R statistic, or lag 1 sample autocorrelation that stopping this early is unsafe!

(e) This example provides evidence against the recommendation of several previous authors that when only a subset of a model's parameters are of interest, convergence of the remaining parameters need not be monitored. While completely unidentified parameters in an overparametrized model may be safely ignored, those that are identified only by the prior must still be monitored, and slow convergence thereof must be remedied. As a result, monitoring a representative subset of all the parameters present is recommended, along with the use of simple reparametrizations where appropriate to speed overall convergence.

Chapter 6

1. Standard hierarchical Bayes calculations yield the following complete conditional distributions:

$$\begin{aligned}
\theta|\boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2, \mathbf{y} &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - F_\theta\boldsymbol{\beta})^T\Sigma_i^{-1}(\mathbf{y} - F_\theta\boldsymbol{\beta})\right\} p(\theta) \\
\boldsymbol{\beta}|\theta, \boldsymbol{\lambda}, \sigma^2, \mathbf{y} &\sim N(Bb, B), \\
\text{where } B^{-1} &= \frac{1}{\sigma^2}F_\theta^T\Sigma_i^{-1}F_\theta + \Sigma_0^{-1} \\
\text{and } b &= \frac{1}{\sigma^2}F_\theta^T\Sigma_i^{-1}\mathbf{y} + \Sigma_0^{-1}\boldsymbol{\beta}_0 \\
\sigma^2|\theta, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y} &\sim IG\left(a_0 + \frac{n}{2},\right. \\
&\left.\left\{b_0^{-1} + \frac{1}{2}\sum_{i=1}^n (y_i - f(\mathbf{x}_i, \theta)\boldsymbol{\beta})^2/\lambda_i\right\}^{-1}\right) .
\end{aligned}$$

Including the λ_i complete conditionals given in Example 6.2, all of these are available in closed form with the exception of that of θ, for which one can use a rejection sampling method (e.g., Metropolis-Hastings).

3. The posterior probability of H_0, $P_G(\theta \in [\theta_L, \theta_U]|x)$, is given by

$$\frac{\int_{\theta_L}^{\theta_U} f(x|\theta)\left[\frac{\pi}{\theta_U - \theta_L}I_{[\theta_L,\theta_U]}(\theta) + (1-\pi)g(\theta)\right]d\theta}{\int f(x|u)\left[\frac{\pi}{\theta_U - \theta_L}I_{[\theta_L,\theta_U]}(u) + (1-\pi)g(u)\right]du} .$$

Setting this expression $\le$ p and keeping in mind that $g(\theta)$ has no support on the interval $[\theta_L, \theta_U]$, we obtain $\mathcal{H}_c$ as the set of all G such that

$$\int f(x|\theta)dG(\theta) \ge c = \frac{1-p}{p}\frac{\pi}{1-\pi}\frac{1}{\theta_U - \theta_L}\int_{\theta_L}^{\theta_U} f(x|\theta)d\theta .$$

4. While the faculty member's solution is mathematically correct, it is not practical since we lack a closed form for the ratio in the expression (i.e. $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ and $p(\boldsymbol{\theta}|\mathbf{y})$ have different and unknown normalizing constants). A conceptually simple alternative would be to run the MCMC algorithm again after deleting y_i, obtaining $\boldsymbol{\theta}^{(g)}$ samples from $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ and subsequeenly obtaining the conditional predictive mean as

$$E(y_i|\mathbf{y}_{(i)}) = \int E(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta}$$

$$\approx \quad \frac{1}{G}\sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)}) \ .$$

A more sophisticated approach (suggested by a more sophisticated faculty member, Prof. Hal Stern) would be to write

$$\begin{aligned}
\int E(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} &= \frac{\int E(y_i|\boldsymbol{\theta})f(\mathbf{y}_{(i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int f(\mathbf{y}_{(i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{\int E(y_i|\boldsymbol{\theta})\frac{f(\mathbf{y}_{(i)}|\boldsymbol{\theta})m(\mathbf{y})}{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\pi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}}{\int \frac{f(\mathbf{y}_{(i)}|\boldsymbol{\theta})m(\mathbf{y})}{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\pi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}} \\
&\approx \frac{\sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)})\frac{f(\mathbf{y}_{(i)}|\boldsymbol{\theta}^{(g)})}{f(\mathbf{y}|\boldsymbol{\theta}^{(g)})}}{\sum_{g=1}^{G} \frac{f(\mathbf{y}_{(i)}|\boldsymbol{\theta}^{(g)})}{f(\mathbf{y}|\boldsymbol{\theta}^{(g)})}}
\end{aligned}$$

where the $\boldsymbol{\theta}^{(g)}$ samples now come directly from the full posterior $p(\boldsymbol{\theta}|\mathbf{y})$, which plays the role of the importance sampling density g in (5.8). In the second line above, the prior terms $\pi(\boldsymbol{\theta})$ cancel in both the numerator and denominator, while the marginal density terms $m(\mathbf{y})$ can be pulled through both integrals and cancelled, since they are free of $\boldsymbol{\theta}$. Note this solution will be accurate provided $p(\boldsymbol{\theta}|\mathbf{y}) \approx p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$, which should hold unless the dataset is quite small or y_i is an extreme outlier.

5. We have

$$\begin{aligned}
f(y_i|\mathbf{y}_{(i)}) &= \int f(y_i|\boldsymbol{\theta},\mathbf{y}_{(i)})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} \\
&= \int f(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} \\
&\approx \int f(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\
&\approx \frac{1}{G}\sum_{g=1}^{G} f(y_i|\boldsymbol{\theta}^{(g)}) \ ,
\end{aligned}$$

where $\boldsymbol{\theta}^{(g)}$ are the MCMC samples from $p(\boldsymbol{\theta}|\mathbf{y})$. Since the CPO is the conditional likelihood, small values indicate poor model fit. Thus we might classify those datapoints having CPO values in the bottom 5-10% as potential outliers meriting further investigation.

6. We have

$$\begin{aligned}
p'_D &= \int P[D(\mathbf{y}^*,\boldsymbol{\theta}) > D(\mathbf{y},\boldsymbol{\theta})|\boldsymbol{\theta}]\, p(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta} \\
&= \int\int I_{\{D(\mathbf{y}^*,\boldsymbol{\theta})>D(\mathbf{y},\boldsymbol{\theta})\}}(\mathbf{y}^*,\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{z})d\mathbf{y}d\boldsymbol{\theta}
\end{aligned}$$

method	$\hat{P}(M=2\|\mathbf{y})$	SD	$\widehat{BF}_{21}$	$\hat{\rho}(1)$	time
CC	.70806	.001721	4848.4	.567	22.8"
RJ-M	.70861	.004058	4861.3	.589	18.7"
RJ-G	.70906	.002394	4871.9	.593	7.9"
RJ-R	.70750	.002004	4835.1	.660	6.7"
Chib-1			4860.7		13.6"
Chib-2			4860.3		14.0"
target	.70865		4862		

Table D.1 *Comparison of different methods, for the simple linear regression example. Here, CC = Carlin and Chib's method; RJ-M = reversible jump using Metropolis steps if the current model is proposed; RJ-G = reversible jump using Gibbs steps if the current model is proposed; RJ-R = reversible jump on the reduced model (i.e., with the regression coefficients integrated out): Chib-1 = Chib's method evaluated at posterior means; and Chib-2 = Chib's method evaluated at frequentist LS solutions.*

$$\approx \frac{1}{G}\sum_{g=1}^{G} I_{\{D(\mathbf{y}^{*(g)},\boldsymbol{\theta}^{(g)})>D(\mathbf{y},\boldsymbol{\theta}^{(g)})\}}(\mathbf{y}^{*(g)}, \boldsymbol{\theta}^{(g)}) .$$

where $(\mathbf{y}^{*(g)}, \boldsymbol{\theta}^{(g)}) \sim p(\mathbf{y}^*, \boldsymbol{\theta}|\mathbf{z}) = f(\mathbf{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{z})$, and I_S denotes the indicator function of the set S. That is,

$$\hat{p}'_D = \frac{\text{number of } D(\mathbf{y}^{*(g)}, \boldsymbol{\theta}^{(g)})s > D(\mathbf{y}, \boldsymbol{\theta}^{(g)})}{G} .$$

The sampler may not have been designed to produce predictive samples $\mathbf{y}^{*(g)}$ but such samples are readily available at iteration g by drawing from $f(\mathbf{y}^*|\boldsymbol{\theta}^{(g)})$, a density that is available in closed form by construction.

8. For each method, we run five independent chains for 50,000 iterations following a 10,000 iteration burn-in period. Table D.1 summarizes our results, reporting the estimated posterior probability for model 2, a batched standard deviation estimate for this probability (using 2500 batches of 100 consecutive iterations), the Bayes factor in favor of model 2, $\widehat{BF}_{21}$, the lag 1 sample autocorrelation for the model indicator, $\hat{\rho}(1)$, and the execution time for the FORTRAN program in seconds. Model 2 is clearly preferred, as indicated by the huge Bayes factor. The "target" probability and Bayes factor are as computed using traditional (non-Monte Carlo) numerical integration by Green and O'Hagan (1998). Programming notes for the various methods are as follows:

(a) For Chib's (1995) marginal density method, we may use a two-block Gibbs sampler that treats σ^2 or τ^2 as θ_1, and $(\alpha, \beta)'$ or $(\gamma, \delta)'$ as $\boldsymbol{\theta}_2$. We experimented with two different evaluation points $(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$; see the table caption.

(b) For Carlin and Chib's (1995) product space search, we have $\boldsymbol{\theta}_1 = (\alpha, \beta, \sigma)$ and $\boldsymbol{\theta}_2 = (\gamma, \delta, \tau)$. For the regression parameters, one may use independent univariate normal pseudopriors that roughly equal the corresponding first-order approximation to the posterior. Acceptable choices are $\alpha|(M=2) \sim N(3000, 52^2)$, $\beta|(M=2) \sim N(185, 12^2)$, $\gamma|(M=1) \sim N(3000, 43^2)$, and $\delta|(M=1) \sim N(185, 9^2)$. For σ^2 and τ^2, taking the pseudopriors equal to the IG priors produces acceptable results. A bit of tinkering reveals that π_1 = .9995, π_2 = .0005 produces $M^{(g)}$ iterates of 1 and 2 in roughly equal proportion.

(c) For Green's (1995) reversible jump algorithm, we use log transforms of the error variances, i.e., $\lambda = \log \sigma^2$, $\omega = \log \tau^2$, to simplify the choice of proposal density. The probabilities of proposing new models are *h(l, l) = h(1, 2) = h(2,1) = h(2, 2)* = 0.5. The dimension matching requirement is automatically satisfied without generating an additional random vector; moreover, the the Jacobian term in the acceptance probability is equal to 1. Since the two regression models are similar in interpretation, when a move between models is proposed, we simply set $(\alpha, \beta, \lambda)' = (\gamma, \delta, \omega)'$ or $(\gamma, \delta, \omega)' = (\alpha, \beta, \lambda)'$. The acceptance probabilities are then given by $\alpha_{1\to 2} = \min\left\{1, \frac{f(\mathbf{y}|\gamma,\delta,\omega,M=2)\pi_2}{f(\mathbf{y}|\alpha,\beta,\lambda,M=1)\pi_1}\right\}$, and $\alpha_{2\to 1} = \min\left\{1, \frac{f(\mathbf{y}|\alpha,\beta,\lambda,M=1)\pi_1}{f(\mathbf{y}|\gamma,\delta,\omega,M=2)\pi_2}\right\}$. When the proposed model is the same as the current model, we update using a standard current-point Metropolis step. That is, for model 1 we draw $(\alpha^*, \beta^*, \lambda^*)' \sim N\left((\alpha^{(k)}, \beta^{(k)}, \lambda^{(k)})', Diag(5000, 250, 1)\right)$ and set $\alpha^{(k+1)}, \beta^{(k+1)}, \lambda^{(k+1)})' = (\alpha^*, \beta^*, \lambda^*)'$ with probability r = $p(\alpha^*, \beta^*, \lambda^*)/p(\alpha^{(k)}, \beta^{(k)}, \lambda^{(k)})$. Similar results hold for model 2. Alternatively, we may perform the "within model" updates using a standard Gibbs step, thus avoiding the log transform and the trivariate normal proposal density, and simply using instead the corresponding full conditional distributions.

Finally, we may also use the reversible jump method on a somewhat reduced model, i.e., we analytically integrate the slopes and intercepts (α, β, γ, and δ) out of the model and use proposal densities for σ^2 and τ^2 that are identical to the corresponding priors. The acceptance probabilities are $\alpha_{1\to 2} = \min\left\{1, \frac{\pi_2}{\pi_1}\frac{f(\mathbf{y}|\tau^2,M=2)}{f(\mathbf{y}|\sigma^2,M=1)}\right\}$ and $\alpha_{2\to 1} = \min\left\{1, \frac{\pi_1}{\pi_2}\frac{f(\mathbf{y}|\sigma^2,M=1)}{f(\mathbf{y}|\tau^2,M=2)}\right\}$.

Summarizing Table D.1, the reversible jump algorithm operating on the reduced model (with the regression coefficients integrated out) appears to be slightly more accurate and faster than the corresponding algorithms operating on the full model. Of course, some extra effort is required to do the integration before programming, and posterior samples are obviously not produced for any parameters no longer appearing in the sampling order. The marginal density (Chib) method does not require preliminary runs (only a point of high posterior density, $\boldsymbol{\theta}'$), and only a rearrangement of existing computer code. The accuracy of results produced by this method is more difficult to assess, but is apparently higher for the given runsize, since repeated runs produced estimates for $\widehat{BF}_{21}$ consistently closer to the target than any of the other methods.

Chapter 7

5. The Bayesian model is

$$p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda}) \propto \prod_{i=1}^{k} \theta_i^{d_i} \exp(-e_i\theta_i) \prod_{i=1}^{k} \theta_i^{\alpha-1}\beta^{-\alpha} \exp(-\theta_i/\beta) I_{S_i}(\theta_i) \; .$$

Hence, using the results of Gelfand, Smith, and Lee (1992), the full conditionals for the θ_i are given by

$$p(\theta_i|\mathbf{y}, \beta, \theta_{j\neq i}) \propto G\left(\theta_i \mid \alpha^*, \ \beta^*\right) I_{(\theta_{i-1}, \theta_{i+1})}(\theta_i) \qquad \text{(D.3)}$$

for $i = 1, \ldots, k$, where $\alpha^* = \alpha + d_i$, $\beta^* = (\beta^{-1} + e_i)^{-1}$, $\theta_0 \equiv 0$, and $\theta_{k+1} \equiv B$. The complete conditional for β is readily available as the untruncated form

$$p(\beta|\mathbf{y}, \boldsymbol{\theta}) = IG\left(a + k\alpha, \ \left\{b^{-1} + \sum_{i=1}^{k} \theta_i\right\}^{-1}\right) \; .$$

The result in the previous problem could be used in generating from the truncated gamma distributions in (D.3), provided subroutines for the gamma cdf (incomplete gamma function) and inverse cdf were available.

6.(b) The simplest solution here is to restrict $\boldsymbol{\theta}$ to lie in an *increasing convex* constraint set,

$$\{\boldsymbol{\theta}: \ \theta_1 > 0 \ , \ \theta_k < B \ , \ 0 < \theta_2 - \theta_1 < \cdots < \theta_k - \theta_{k-1}\} \; .$$

However, Gelman (1996) points out that this is a very strong set of constraints, and one which becomes stronger as k increases (e.g., as we graduate more years of data, or even subdivide our yearly data into monthly intervals). It is akin to fitting a quadratic to the observed data.

9.(a) This is the approach taken by Carlin and Polson (1992).

(b) This is the approach taken by Albert and Chib (1993a) and (in a multivariate response setting) by Cowles et al. (1996).

10.(a) Running 3 chains of our Gibbs-Metropolis sampler for 1000 iterations produced the following posterior medians and 95% equal-tail credible intervals: for ϕ, 1.58 and (0.28, 4.12); for $X(\mathbf{s}_A)$, .085 and (.028, .239); and for $X(\mathbf{s}_B)$, .077 and (.021, .247). The spatially smoothed predictions at points A and B are consistent with the high values at monitoring stations 4 and 5 to the east, the moderate value at station 3 to the west, and the low value at station 8 to the far south.

(b) Ozone is a wind-borne pollutant, so if a prevailing wind (say, from west to east) dominated Atlanta's summer weather pattern, it would be reasonable to expect some anisotropy (in this hypothetical case, more spatial similarity west to east than north to south).

A credible interval for β_{12} that excluded 0 would provide evidence that the anisotropic model is warranted in this case. One might also compare the two models formally (assuming proper priors) via Bayes factors (see Subsections 6.3 and 6.4) or informally via predictive or penalized likelihood criteria (see Subsection 6.5).

11. Based on 1000 Gibbs samples from single chain generated in the **BUGS** language (and collected after a 500-iteration burn-in period), Spiegelhalter et al. (1995b) report the following posterior mean and standard error estimates:

(a) α, $.73 \pm .14$; κ_θ, $-.54 \pm .16$. Thus, the observed rates are slightly lower than expected from the standard table, but higher for those counties having a greater percentage of their population engaged in agriculture, fishing, and forestry (AFF).

(b) α, $.39 \pm .12$ (an overall intercept term cannot be fit in the presence of our translation-invariant CAR prior). While the covariate remains significant in this model, its reduced importance would appear due to this model's ability to account for spatial similarity. (We would expect some confounding of these two effects, since counties where AFF employment is high would often be adjacent.)

(c) Spiegelhalter et al. (1998) obtain p_D = 39.7, *DIC* = 101.9 for the part (a) model (covariate plus exchangeble random effects), and p_D = 29.4, *DIC* = 89.0 for the part (b) model (covariate plus spatial random effects). Thus, *DIC* prefers the spatial model, largely due to its slightly smaller effective sample size.

(d) The fact that the spatial plus covariate model fits slightly better than the heterogeneity plus covariate model suggests that there is some excess spatial variability in the data that is not explained by the covariate. To get a better idea of this variability, we could map the fitted values $\hat{\phi}_i = E(\phi_i|\mathbf{y})$, thinking of them as spatial residuals. A pattern in this map might then indicate the presence of a lingering spatially-varying covariate not yet accounted for.

(e) Suppose we begin with proper hyperpriors for τ and λ. The translation invariance of the CAR prior forces to set $\kappa_\theta = 0$. Alternatively, we could estimate an overall level κ_θ after imposing the constraint $\sum_i \theta_i = 0$ and/or $\sum_i \phi_i = 0$.

Appendix B

1. This proof is similar to that given in Example B.5.
2. For action a, the posterior risk is

$$\begin{aligned} \rho(\pi, a) &= \sum_i l(\theta_i, a)\pi(\theta_i|\mathbf{x}) \\ &= 1 - \pi(a|\mathbf{x}) , \end{aligned}$$

since $l(\theta_i, a) = 1$ in the sum unless $\theta_i = a$, and $\sum_i \pi(\theta_i|\mathbf{x}) = 1$. To minimize this quantity, we take a to maximize $\pi(a|\mathbf{x})$, hence the Bayes action is the posterior mode.

9. The optimal value of c is

$$c = \frac{(\rho\beta)^2}{\sigma^2 + (\rho\beta)^2} .$$

Because $c \leq 1$, the optimal estimator shrinks the unbiased estimator toward 0 by an amount that depends on the relative size of σ and $(\rho\beta)$. As $\sigma \to \infty$, $c \to 0$ (variance reduction is all important). When $\sigma \to 0$, $c \to 1$ (when the variance is small, bias reduction is the dominant goal).

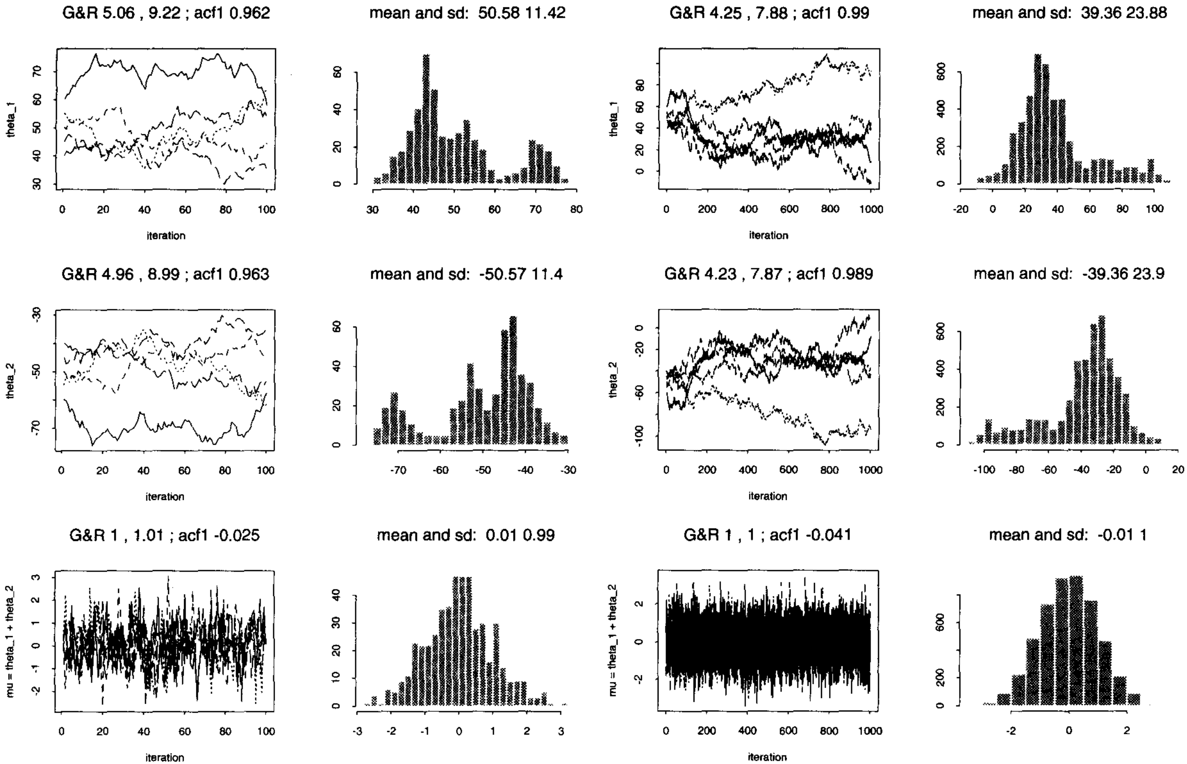


Figure D.1 *MCMC analysis of $N(\theta_1 + \theta_2, 1)$ likelihood, Case 1: both parameters have prior mean 50 and prior standard deviation 1000.*

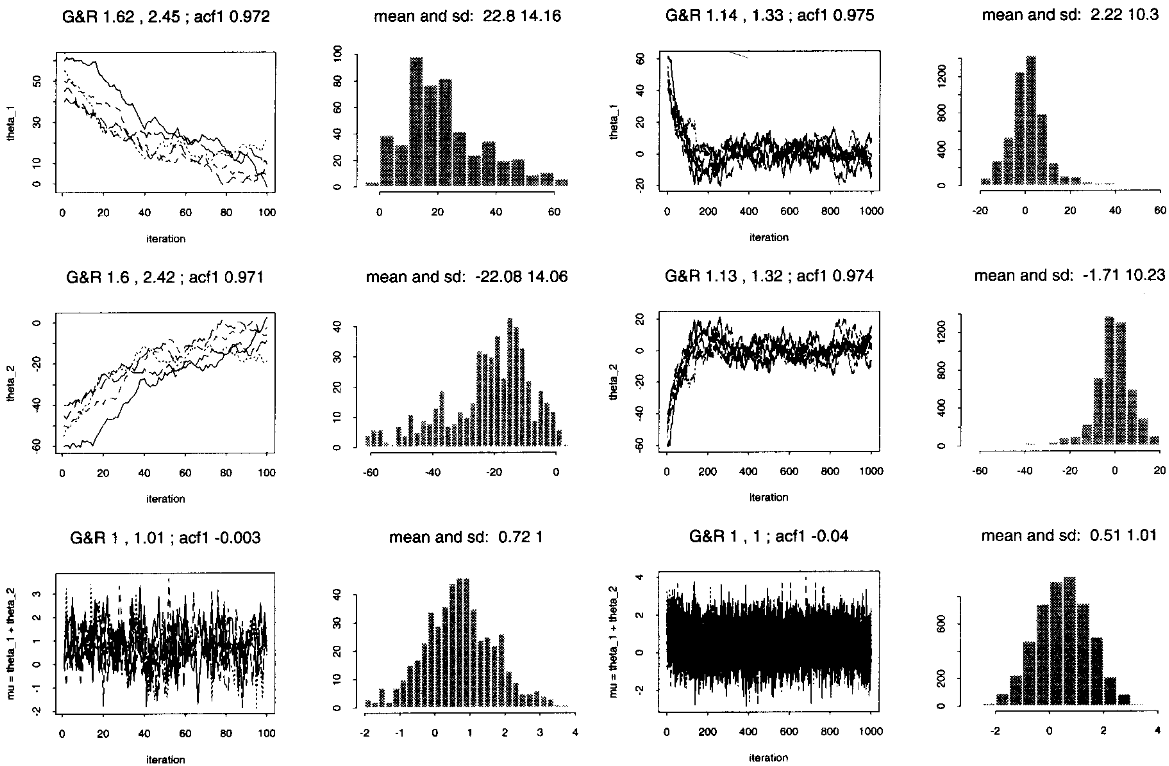


Figure D.2 *MCMC analysis of N($\theta_1 + \theta_2$,1) likelihood, Case 2: both parameters have prior mean 50 and prior standard deviation 10.*

References

Abrams, D.I., Goldman, A.I., Launer, C., Korvick, J.A., Neaton, J.D., Crane, L.R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., Cohn, D.L., Harris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J.H., Thompson, M., Deyton, L., and the Terry Beirn Community Programs for Clinical Research on AIDS (1994). Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant of or have failed zidovudine therapy. *New England Journal of Medicine*, **330**, 657-662.

Adler, S.L (1981). Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions. Physical Review D, **23**, 2901-2904.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.

Albert, J.H. (1993). Teaching Bayesian statistics using sampling methods and MINITAB. *The American Statistician*, 47, 182-191.

Albert, J.H. (1996). *Bayesian Computation Using Minitab*. Belmont, CA: Wadsworth.

Albert, J.H. and Chib, S. (1993a). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, 88, 669-679.

Albert, J.H. and Chib, S. (1993b). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business* and Economic Statistics, 11, 1-15.

Almond, R.G. (1995). *Graphical Belief Modeling.* London: Chapman and Hall.

Ames, B.N., Lee, F.D., and Durston, W.E. (1973). An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proceedings of the National Academy of Sciences*, 70, 782-786.

Andrews, D.F. and Mallows, C.L. (1974). Scale mixtures of normality. *J. Roy. Statist. Soc., Ser. B,* 36, 99-102.

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *Ann. Statist.*, 2, 1152-1174.

Bates, D.M. and Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*. New York: Wiley.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc. London,* **53**, 370-418. Reprinted, with an introduction by George Barnard, in 1958 in *Biometrika,* 45, 293-315.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Berger, J.O. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness in Bayesian Statistics*, ed. J. Kadane, Amsterdam: North Holland.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.

Berger, J.O. (1994). An overview of robust Bayesian analysis (with discussion). *Test*, **3**, 5-124.

Berger, J.O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. To appear, *J. Amer. Statist. Assoc.*

Berger, J.O. and Berliner, L.M. (1986). Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *Annals of Statistics*, **14**, 461-486.

Berger, J.O. and Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, **76**, 159-165.

Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statistical Science*, **2**, 317-52.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 25-44.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence (with discussion). *J. Am. Statist. Assoc.*, **82**, 112-122.

Berger, J.O. and Wolpert, R. (1984). *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics Monograph Series.

Berger, R.L. and Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets (with discussion). *Statistical Science*, **11**, 283-319.

Bernardinelli, L., Clayton, D.G., and Montomoli, C. (1995). Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine*, **14**, 2411-2431.

Bernardinelli, L., Clayton, D.G., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, **14**, 2433-2443.

Bernardinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, **11**, 903-1007.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. B*, **41**, 113-147.

Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M., eds. (1992). *Bayesian Statistics 4*. Oxford: Oxford University Press.

Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M., eds. (1996). *Bayesian Statistics 5*. Oxford: Oxford University Press.

Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M., eds. (1999). *Bayesian Statistics 6*. Oxford: Oxford University Press.

Bernardo, J.M., DeGroot, M.H., Lindley, D.V., and Smith, A.F.M.; eds. (1980). *Bayesian Statistics*. Valencia: University Press.

Bernardo, J.M., DeGroot, M.H., Lindley, D.V., and Smith, A.F.M., eds. (1985). *Bayesian Statistics 2*. Amsterdam: North Holland.

Bernardo, J.M., DeGroot, M.H., Lindley, D.V., and Smith, A.F.M, eds. (1988).

Bayesian Statistics 3. Oxford: Oxford University Press.

Bernardo, J.M. and Smith, A.F.M. (1994). Bayesian Theory. New York: Wiley.

Berry, D.A. (1991). Bayesian methods in phase III trials. *Drug Information Journal*, **25**, 345-368.

Berry, D.A. (1993). A case for Bayesianism in clinical trials (with discussion). *Statistics in Medicine*, **12**, 1377-1404.

Berry, D.A. (1996). *Statistics: A Bayesian Perspective*. Belmont, CA: Duxbury.

Berry, D.A., Müller, P., Grieve, A.P., Smith, M., Parke, T., Blazek, R., Mitchard, N., Brearly, C., and Krams, M. (2000). Bayesian designs for dose-ranging drug trials (with discussion). To appear in *Case Studies in Bayesian Statistics, Volume V*, C. Gatsonis et al., eds., New York: Springer-Verlag.

Berry, D.A. and Stangl, D.K., eds. (1996). *Bayesian Biostatistics*. New York: Marcel Dekker.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **36**, 192-236.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc., Ser. B*, **48**, 259-302.

Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc., Ser. B*, **55**, 25-37.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3-66.

Besag, J., York, J.C., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Best, N.G., Cowles, M.K., and Vines, K. (1995). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Best, N.G., Ickstadt, K., Wolpert, R.L., Cockings, S., Elliott, P., Bennett, J., Bottle, A., and Reed, S. (2000). Modeling the impact of traffic-related air pollution on childhood respiratory illness (with discussion). To appear in Case *Studies in Bayesian Statistics, Volume V*, C. Gatsonis et al., eds., New York: Springer-Verlag.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.*, **57**, 269-326.

Bliss, C.I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, **22**, 134-167.

Box, G.E.P. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. London: Addison-Wesley.

Box, G.E.P. and Muller, M.E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.*, **29**, 610-611.

Brandwein, A.C. and Strawderman, W.E. (1990). Stein estimation: The spherically symmetric case. *Statistical Science*, **5**, 356-369.

Breslow, N.E. (1984). Extra-Poisson variation in log-linear models. *Appl. Statist.*, **3**,38-44.

Breslow, N.E. (1990). Biostatistics and Hayes (with discussion). *Statistical Science*, **5**, 269-298.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9-25.

Broffitt, J.D. (1988). Increasing and increasing convex Bayesian graduation (with discussion). *Trans. Soc. Actuaries,* **40**, 115-148.

Brooks, S.P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Statist.,* **7**, 434-455.

Brooks, S.P. and Roberts, G.O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing,* **8**, 319-335.

Brown, L.D. (2000). An essay on statistical decision theory. To appear, *J. Amer. Statist. Assoc.*

Burden, R.L., Faires, J.D., and Reynolds, A.C. (1981). *Numerical Analysis,* 2nd ed. Boston: Prindle, Weber and Schmidt.

Butler, S.M. and Louis, T.A. (1992). Random effects models with non-parametric priors. *Statistics in Medicine,* **11**, 1981-2000.

Carlin, B.P. (1992). A simple Monte Carlo approach to Bayesian graduation. *Trans. Soc. Actuaries,* **44**, 55-76.

Carlin, B.P. (1996). Hierarchical longitudinal modeling. In *Markov Chain Monte Carlo in Practice*, eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, London: Chapman and Hall, pp. 303-319.

Carlin, B.P., Chaloner, K., Church, T., Louis, T.A., and Matts, J.P. (1993). Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *The Statistician,* **42**, 355-367.

Carlin, B.P., Chaloner, K.M., Louis, T.A., and Rhame, F.S. (1995). Elicitation, monitoring, and analysis for an AIDS clinical trial (with discussion). In *Case Studies in Bayesian Statistics, Volume II,* eds. C. Gatsonis, J.S. Hodges, R.E. Kass and N.D. Singpurwalla, New York: Springer-Verlag, pp. 48-89.

Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc., Ser. B,* **57**, 473-484.

Carlin, B.P. and Gelfand, A. E. (1990). Approaches for empirical Bayes confidence intervals. *J. Amer. Statist. Assoc.,* **85**, 105-114.

Carlin, B.P. and Gelfand, A.E. (1991a). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J. Roy. Statist. Soc., Ser. B,* **53**, 189-200.

Carlin, B.P. and Gelfand, A.E. (1991b). An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statistics and Computing,* **1**, 119-128.

Carlin, B.P., Kadane, J.B., and Gelfand, A.E. (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics* , **54**, 964-975.

Carlin, B.P., Kass, R.E., Lerch, F.J., and Huguenard, B.R. (1992). Predicting working memory failure: A subjective Bayesian approach to model selection. *J. Amer. Statist. Assoc.,* **87**, 319-327.

Carlin, B.P. and Klugman, S.A. (1993). Hierarchical Bayesian Whittaker graduation. *Scandinavian Actuarial Journal,* **1993.2** , 183-196.

Carlin, B.P. and Louis, T.A. (1996). Identifying prior distributions that produce specific decisions, with application to monitoring clinical trials. In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner,* eds. D. Berry, K. Chaloner, and J. Geweke, New York: Wiley, pp. 493-503.

Carlin, B.P. and Louis, T.A. (2000). Empirical Bayes: past, present and future.

To appear, *J. Amer. Statist. Assoc.*

Carlin, B.P. and Polson, N.G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canad. J. Statist.*, **19**, 399-405.

Carlin, B.P. and Polson, N.G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 577-586.

Carlin, B.P., Polson, N.G., and Stoffer, D.S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Amer. Statist. Assoc.*, **87**, 493-500.

Carlin, B.P. and Sargent, D.J. (1996). Robust Bayesian approaches for clinical trial monitoring. *Statistics in Medicine,* **15**, 1093-1106.

Carlin, J.B. and Louis, T.A. (1985). Controlling error rates by using conditional expected power to select tumor sites. *Proc. Biopharmaceutical Section of the Amer. Statist. Assoc.*, 11-18.

Carter, C.K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika,* **81**, 541-553.

Carter, C.K. and Kohn, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika,* **83**, 589-601.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, **39**, 83-87.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Pacific Grove, CA: Wadsworth & Brooks-Cole.

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.

Casella, G. and Hwang, J. (1983). Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *J. Amer. Statist. Assoc.*, **78**, 688-698.

Centers for Disease Control and Prevention, National Center for Health Statistics (1988). *Public Use Data Tape Documentation Compressed Mortality File, 1968-1985*. Hyattsvile, Maryland: U.S. Department of Health and Human Services.

Chaloner, K. (1996). The elicitation of prior distributions. In *Bayesian Biostatistics*, eds. D. Berry and D. Stangl, New York: Marcel Dekker, pp. 141-156.

Chaloner, K., Church, T., Louis, T.A., and Matts, J.P. (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician*, **42**, 341-353.

Chen, D.-G., Carter, E.M., Hubert, J.J., and Kim, P.T. (1999). Empirical Bayes estimation for combinations of multivariate bioassays. *Biometrics*, **55**, 1038-1043.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.*, **89**, 818-824.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.*, **90**, 1313-1321.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**, 79-97.

Chib, S. and Carlin, B.P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*, **9**, 17-26.

Chib, S. and Greenberg, E. (1995a). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327-335.

Chib, S. and Greenberg, E. (1995b). Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models. *Journal of Econometrics*, **68**, 339-360.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, 347-361.

Choi, S., Lagakos, S.W., Schooley, R.T., and Volberding, P.A. (1993) CD4+ lymphocytes are an incomplete surrogate marker for clinical progession in persons with asymptomatic HIV infection taking zidovudine. *Ann. Internal Med.*, **118**, 674-680.

Christiansen, C.L. and Morris, C.N. (1997). Hierarchical Poisson regression modeling. *J. Amer. Statist. Assoc.*, **92**, 618-632.

Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**, 467-485.

Clayton, D.G. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English, and R. Stern, eds., Oxford: Oxford University Press.

Clayton, D.G. and Kaldor, J.M. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671-681.

Clyde, M., DeSimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.*, **91**, 1197-1208.

Concorde Coordinating Committee (1994). Concorde: MRC/ANRS randomized double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet*, **343**, 871-881.

Conlon, E.M. and Louis, T.A. (1999). Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In *Disease Mapping and Risk Assessment for Public Health*, A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, and R. Bertollini, eds., New York: Wiley, pp. 31-47.

Cooper, H. and Hedges, L. eds. (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Cornfield, J. (1966a). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, **20**, 18-23.

Cornfield, J. (1966b). A Bayesian test of some classical hypotheses - with applications to sequential clinical trials. *J. Amer. Statist. Assoc.*, **61**, 577-594.

Cornfield, J. (1969). The Bayesian outlook and its applications. *Biometrics*, **25**, 617-657.

Cowles, M.K. (1994). Practical issues in Gibbs sampler implementation with application to Bayesian hierarchical modeling of clinical trial data. Unpublished Ph.D. disssertation, Division of Biostatistics, University of Minnesota.

Cowles, M.K. and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.*, **91**, 883-904.

Cowles, M.K., Carlin, B.P., and Connett, J.E. (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *J. Amer. Statist. Assoc.*, **91**, 86-98.

Cowles, M.K. and Rosenthal, J.S. (1998). A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. *Statistics and Computing*, **8**, 115-124.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Creasy, M.A. (1954). Limits for the ratio of means. *J. Roy. Statist. Soc., Ser. B*, **16**, 186-194.

Cressie, N. (1989). Empirical Bayes estimates of undercount in the decennial census. *J. Amer. Statist. Assoc.*, **84**, 1033-1044.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, revised ed. New York: Wiley.

Cressie, N. and Chan, N.H. (1989). Spatial modeling of regional variables. *J. Amer. Statist. Assoc.*, **84**, 393-401.

Damien, P., Laud, P.W., and Smith, A.F.M. (1995). Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *J. Roy. Statist. Soc., Ser. B*, **57**, 547-563.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Roy. Statist. Soc., Ser. B*, **61**, 331-344.

Deely, J.J. and Lindley, D.V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.*, **76**, 833-841.

DeGroot, M.H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Dellaportas, P. and Smith, A.F.M. (1993). Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. *J. Roy. Statist. Assoc., Ser. C (Applied Statistics)*, **42**, 443-459.

Dellaportas, P., Forster, J.J., and Ntzoufras, 1. (1998). On Bayesian model and variable selection using MCMC. Technical report, Department of Statistics, Athens University of Economics and Business.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc., Ser. B*, **39**, 1-38.

Dempster, A.P., Selwyn, M.R., and Weeks, B.J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.*, **78**, 221-227.

DerSimonian, R. and Laird, N.M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177-188.

DeSantis, F. and Spezzaferri, F. (1999). Methods for default and robust Bayesian model comparison: the fractional Bayes factor approach. *International Statistical Review*, **67**, 267-286.

DeSouza, C.M. (1991). An empirical Bayes formulation of cohort models in cancer epidemiology. *Statistics in Medicine*, **10**, 1241-1256.

Devine, O.J. (1992). Empirical Bayes and constrained empirical Bayes methods for estimating incidence rates in spatially aligned areas. Unpublished Ph.D. dissertation, Division of Biostatistics, Emory University.

Devine, O.J., Halloran, M.E., and Louis, T.A. (1994). Empirical Bayes methods for stabilizing incidence rates prior to mapping. *Epidemiology*, **5**, 622-630.

Devine, O.J. and Louis, T.A. (1994). A constrained empirical Bayes estimator for incidence rates in areas with small populations. *Statistics in Medicine,* **13**, 1119-1133.

Devine, O.J., Louis, T.A., and Halloran, M.E. (1994). Empirical Bayes estimators for spatially correlated incidence rates. *Environmetrics,* **5**, 381-398.

Devine, O.J., Louis, T.A., and Halloran, M.E. (1996). Identifying areas with elevated incidence rates using empirical Bayes estimators. *Geographical Analysis,* **28**, 187-199.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

Dey, D.K. and Berger, J.O. (1983). On truncation of shrinkage estimators in simultaneous estimation of normal means. *J. Amer. Statist. Assoc.,* **78**, 865-869.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc., Ser. B,* **56**, 363-375.

Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns.* London: Academic Press.

Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *J. Roy. Statist. Soc., Ser. C,* **47**, 299-350.

Doss, H. and Narasimhan, B. (1994). Bayesian Poisson regression using the Gibbs sampler: Sensitivity analysis through dynamic graphics. Technical report, Department of Statistics, The Ohio State University.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc., Ser. B,* **57**, 45-97.

DuMouchel, W.H. (1988). A Bayesian model and a graphical elicitation procedure for multiple comparisons. In *Bayesian Statistics 3,* eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Oxford: Oxford University Press, pp. 127-145.

DuMouchel, W.H. and Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.,* **78**, 293-315.

Eberly, L.E. and Carlin, B.P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. To appear, *Statistics in Medicine.*

Ecker, M.D. and Gelfand, A.E. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology,* **31**, 67-83.

Edwards, R.G. and Sokal, A.D. (1988). Generalization of the Fortium-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review D,* **38**, 2009-2012.

Edwards, W., Lindman, H., and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psych. Rev.,* **70**, 193-242.

Efron, B. (1986). Why isn't everyone a Bayesian? (with discussion). *Amer. Statistician,* **40**, *1-11.*

Efron, B. (1996). Empirical Bayes methods for combining likelihoods (with discussion). *J. Amer. Statist. Assoc.,* **91**, 538-565.

Efron, B. and Morris, C.N. (1971). Limiting the risk of Bayes and empirical Bayes estimators - Part 1: The Bayes case. *J. Amer. Statist. Assoc.,* **66**, 807-815.

Efron, B. and Morris, C.N. (1972a). Limiting the risk of Bayes and empirical Bayes estimators - Part 11: The empirical Bayes case. *J. Amer. Statist. Assoc.*, **67**, 130-139.

Efron, B. and Morris, C.N. (1972b). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika,* **59**, 335-347.

Efron, B. and Morris, C. (1973a). Stein's estimation rule and its competitors - An empirical Bayes approach. *J. Amer. Statist. Assoc.*, **68**, 117-130.

Efron, B. and Morris, C. (1973b). Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **35**, 379-421.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, **70**, 311-319.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American,* **236**, 119-127.

Erkanli, A., Soyer, R., and Costello, E.J. (1999). Bayesian inference for prevalence in longitudinal two-phase studies. *Biometrics*, **55**, 1145-1150.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.,* **89**, 268-277.

Escobar, M.D. and West, M. (1992). Computing Bayesian nonparametric hierarchical models. Discussion Paper 92-A20, Institute for Statistics and Decision Sciences, Duke University.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, **90**, 577-588.

Etzioni, R. and Carlin, B.P. (1993). Bayesian analysis of the Ames Salmonella/microsome assay. In *Case Studies in Bayesian Statistics (Lecture Notes in Statistics, Vol. 83)*, eds. C. Gatsonis, J.S. Hodges, R.E. Kass, and N.D. Singpurwalla, New York: Springer-Verlag, pp. 311-323.

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, **10**, 254-272; discussion and rejoinder: 11, 54-64.

Fay R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Ferguson. T.S. (1967). *Mathematical Statistics: A Decision Theory Approach.* New York: Academic Press.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209-230.

Fieller, E.C. (1954). Some problems in interval estimation. *J. Roy. Statist. Soc., Ser. B*, **16**, 175-185.

Finney, D.J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika,* **34**, 320-334.

Fleming, T.R. (1992). Evaluating therapeutic interventions: Some issues and experiences (with discussion). *Statistical Science,* **7**, 428-456.

Freedman, L.S. and Spiegelhalter, D.J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician,* **33**, 153-160.

Freedman, L.S. and Spiegelhalter, D.J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Tri-*

als, **10**, 357-367.

Freedman, L.S. and Spiegelhalter, D.J. (1992). Application of Bayesian statistics to decision making during a clinical trial. *Statistics in Medicine*, **11**, 23-35.

Fuller, W.A. (1987). *Measurement Error Models*. New York: Wiley.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. London: Chapman and Hall.

Gatsonis, C., Hodges, J.S., Kass, R.E., and Singpurwalla, N.D., eds. (1993). *Case Studies in Bayesian Statistics*. New York: Springer-Verlag.

Gatsonis, C., Hodges, J.S., Kass, R.E., and Singpurwalla, N.D., eds. (1995). *Case Studies in Bayesian Statistics, Volume II*. New York: Springer-Verlag.

Gatsonis, C., Hodges, J.S., Kass, R.E., McCulloch, R.E., Rossi, P., and Singpurwalla, N.D., eds. (1997). *Case Studies in Bayesian Statistics, Volume III*. New York: Springer-Verlag.

Gatsonis, C., Kass, R.E., Carlin, B.P., Carriquiry, A.L., Gelman, A., Verdinelli, I., and West, M., eds. (1999). *Case Studies in Bayesian Statistics, Volume IV*. New York: Springer-Verlag.

Gatsonis, C., Kass, R.E., Carlin, B.P., Carriquiry, A.L., Gelman, A., Verdinelli, I., and West, M., eds. (2000). *Case Studies in Bayesian Statistics, Volume V*. New York: Springer-Verlag.

Gaver, D.P. and O'Muircheartaigh, I.G. (1987). Robust empirical Bayes analyses of event rates. *Technometrics*, **29**, 1-15.

Geisser, S. (1993). *Predictive Inference*. London: Chapman and Hall.

Gelfand, A.E. and Carlin, B.P. (1993). Maximum likelihood estimation for constrained or missing data models. *Canad. J. Statist.*, **21**, 303-311.

Gelfand, A.E., Carlin, B.P., and Trevisani, M. (2000). On computation using Gibbs sampling for multilevel models. Technical report, Department of Statistics, University of Connecticut.

Gelfand, A.E. and Dalal, S.R. (1990). A note on overdispersed exponential families. *Biometrika*, **77**, 55-64.

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc., Ser. B*, **56**, 501-514.

Gelfand, A.E., Dey, D.K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 147-167.

Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1-11.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.*, **85**, 972-985.

Gelfand, A.E., and Lee, T.-M. (1993). Discussion of the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods. *J. Roy. Statist. Soc., Ser. B*, **55**, 72-73.

Gelfand, A.E. and Sahu, S.K. (1994). On Markov chain Monte Carlo acceleration. *J. Comp. Graph. Statist.*, **3**, 261-276.

Gelfand, A.E. and Sahu, S.K. (1999). Identifiability, improper priors and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.*, **94**, 247-253.

Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1995). Efficient parametrizations for normal linear mixed models. *Biometrika*, **82**, 479-488.

Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1996). Efficient parametrizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press, pp. 165-180.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85, 398-409.

Gelfand, A.E., Smith, A.F.M., and Lee, T-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.*, **87**, 523-532.

Gelman, A. (1996). Bayesian model-building by pure thought: some principles and examples. *Statistica Sinica*, **6**, 215-232.

Gelman, A., Carlin, J., Stern, H., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gelman, A., Meng, X.-L., and Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733-807.

Gelman, A., Roberts, G.O., and Gilks, W.R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 599-607.

Gelman. A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457-511.

Geman. S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

Gentle. J.E. (1998). *Random Number Generation and Monte Carlo Methods*. New York: Springer-Verlag.

Genz. A. and Kass, R. (1997). Subregion adaptive integration of functions having a dominant peak. *J. Comp. Graph. Statist.*, **6**, 92-111.

George. E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **88**, 881-889.

Geweke. J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317-1339.

Geweke. J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In *Bayesian Statistics 4*, J.M. Bernardo. J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 169-193.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science*. **7**. 473-511.

Geyer, C.J. (1993). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical report No. 568, School of Statistics, University of Minnesota.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc., Ser. B*, **54**, 657-699.

Geyer, C.J. and Thompson, E.A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.,* **90**, 909-920.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *J. Amer. Statist. Assoc.,* **87**, 533-540.

Ghosh, M. and Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. J. *Amer. Statist. Assoc.,* **81**, 1058-1062.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science,* **9**, 55-93.

Gilks, W.R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics 4,* J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 641-649.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D., and Kirby, A.J. (1993). Modelling complexity: Applications of Gibbs sampling in medicine. *J. Roy. Statist. Soc., Ser. B,* **55**, 38-52.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Gilks, W.R. and Roberts, G.O. (1996). Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice,* W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, D.J., eds, London: Chapman and Hall, pp. 89-114.

Gilks, W.R., Roberts, G.O., and Sahu, S.K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Stat. Assoc.,* **93**, 1045-1054.

Gilks, W.R., Thomas, A., and Spiegelhalter, D.J. (1992) Software for the Gibbs sampler. *Computing Science and Statistics,* **24**, 439-448.

Gilks, W.R., Wang, C.C., Yvonnet, B., and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics,* **49**, *441-453.*

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. Roy. Statist. Soc., Ser. C (Applied Statistics),* **41**, 337-348.

Godsill, S.J. (2000). On the relationship between MCMC model uncertainty methods. Technical Report CUED /F-INFENG/TR.305, Engineering Department, Cambridge University.

Goel, P.K. (1988). Software for Bayesian analysis: Current status and additional needs (with discussion). In *Bayesian Statistics 3,* J. M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 173-188.

Goldman, A.I., Carlin, B.P., Crane, L.R., Launer, C., Korvick, J.A., Deyton, L., and Abrams, D.I. (1996). Response of CD4+ and clinical consequences to treatment using ddI or ddC in patients with advanced HIV infection. J. *Acquired Immune Deficiency Syndromes and Human Retrovirology,* **11**, 161-169.

Goldstein, H. (1995). *Kendall's Library of Statistics 3: Multilevel Statistical Models,* 2nd ed. London: Arnold.

Goldstein, H. and Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. Roy. Statist. Soc., Ser. A,* **159**, 385-443.

Goldstein, M. (1981). Revising previsions: A geometric interpretation. *J. Roy. Statist. Soc., Ser. B,* **43**, 105-130.

Good, I.J. (1965). *The Estimation of Probabilities.* Cambdrige, MA: MIT Press.

Grambsch, P. and Therneau, T.M. (1994). Proportional hazards tests and diag-

nostics based on weighted residuals. *Biometrika,* **81**, 515-526.
Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika,* **82**, 711-732.
Green, P.J. and Murdoch, D.J. (1999). Exact sampling for Bayesian inference: towards general purpose algorithms (with discussion). In *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 301-321.
Green, P.J. and O'Hagan, A. (1998). Carlin and Chib do not need to sample from pseudopriors. Research Report 98-1, Department of Statistics, University of Nottingham.
Greenhouse, J.B. and Wasserman, L.A. (1995). Robust Bayesian methods for monitoring clinical trials. *Statistics in Medicine,* **14**, 1379-1391.
Gustafson, P. (1994). Hierarchical Bayesian analysis of clustered survival data. Technical Report No. 144, Department of Statistics, University of British Columbia.
Guttman, I. (1982). *Linear Models: An Introduction.* New York: Wiley.
Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods.* London: Methuen.
Han, C. and Carlin, B.P. (2000). MCMC methods for computing Bayes factors: A comparative review. Research Report 2000-01, Division of Biostatistics, University of Minnesota, 2000.
Handcock, M.S. and Stein, M.L. (1993). A Bayesian analysis of kriging. *Technometrics,* **35**, 403-410.
Handcock, M.S. and Wallis, J.R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Amer. Statist. Assoc.,* **89**, 368-390.
Hartigan, J.A. (1983). *Bayes Theory*. New York: Springer-Verlag.
Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* **57**, 97-109.
Heidelberger, P. and Welch, P.D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research,* **31**, 1109-1144.
Higdon, D.M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.,* **93**, 585-595.
Hill, J.R. (1990). A general framework for model-based statistics. *Biometrika,* **77**, 115-126.
Hills, S.E. and Smith, A.F.M. (1992). Parametrization issues in Bayesian inference. In *Bayesian Statistics 4,* eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, pp. 641-649, Oxford: Oxford University Press.
Hobert, J.P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.,* **91**, 1461-1473.
Hodges, J.S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *J. Roy. Statist. Soc., Series B*, **60**, 497-536.
Hodges, J.S. and Sargent, D.J. (1998). Counting degrees of freedom in hierarchical and other richly- par ameterised models. Technical report, Division of Biostatistics, University of Minnesota.
Hogg, R.V. and Craig, A.T. (1978). *Introduction to Mathematical Statistics*, 4th

ed. New York: Macmillan.

Hui, S.L. and Berger, J.O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *J. Amer. Statist. Assoc.*, **78**, 753-759.

Hurn, M. (1997). Difficulties in the use of auxiliary variables in Markov chain Monte carlo methods. *Statistics and Computing*, **7**, 35-44.

Jacobson, M.A., Besch, C.L., Child, C., Hafner, R., Matts, J.P., Muth, K., Wentworth, D.N., Neaton, J.D., Abrams, D., Rimland, D., Perez, G., Grant, I.H., Saravolatz, L.D., Brown, L.S., Deyton, L., and the Terry Beirn Community Programs for Clinical Research on AIDS (1994). Primary prophylaxis with pyrimethamine for toxoplasmic encephalitis in patients with advanced human immunodeficiency virus disease: Results of a randomized trial. J. *Infectious Diseases*, **169**, 384-394.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, **1**, Berkeley, CA: Univ. of California Press, pp. 361-379.

Jeffreys, H. (1961). *Theory of Probability,* 3rd ed. Oxford: University Press.

Johnson, V.E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *J. Amer. Statist. Assoc.*, **93**, 238-248.

Kackar, R.N. and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.*, **79**, 853-862.

Kadane, J.B. (1986). Progress toward a more ethical method for clinical trials. *J. of Medical Philosophy,* **11,** 385-404.

Kadane, J.B., Chan, N.H., and Wolfson, L.J. (1996). Priors for unit root models. *J. of Econometrics,* **75**, 99-111.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters. S.C. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.*, **75**, 845-854.

Kadane, J.B. and Wolfson, L.J. (1996). Priors for the design and analysis of clinical trials. In *Bayesian Biostatistics*, eds. D. Berry and D. Stangl, New York: Marcel Dekker, pp. 157-184.

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *J. Roy. Statist. Soc., Ser B*, **40**, 214-221.

Kaluzny, S.P., Vega, S.C., Cardoso, T.P., and Shelly, A.A. (1998). *S+SpatialStats: User's Manual for Windows and UNIX*. New York: Springer-Verlag.

Kass, R.E., Carlin, B.P., Gelman, A., and Neal, R. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician,* **52**, 93-100.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773-795.

Kass, R.E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.*, **84**, 717-726.

Kass, R.E., Tierney, L., and Kadane, J. (1988). Asymptotics in Bayesian computation (with discussion). In *Bayesian Statistics 3*, J.M. Bernardo, M.H. DeGroot, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 261-278.

Kass, R.E., Tierney, L., and Kadane, J. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika,* **76**, 663-674.

Kass, R.E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. Roy. Statist. Soc. B,* **54**, 129-144.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.*, **91**, 1343-1370.

Killough, G.G., Case, MA., Meyer, K.R., Moore, R.E., Rope, S.K., Schmidt, D.W., Schleien, B., Sinclair, W.K., Voillequé, P.G., and Till, J.E. (1996). Task 6: Radiation doses and risk to residents from FMPC operations from 1951-1988. Draft report, Radiological Assessments Corporation, Neeses, SC.

Kolassa, J.E. and Tanner, M.A. (1994). Approximate conditional inference in exponential families via the Gibbs sampler. *J. Amer. Statist. Assoc.,* **89**, 697-702.

Knorr-Held, L. and Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics,* **56**, 13-21.

Kreft, I.G.G., de Leeuw, J., and van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, and VARCL. *The American Statistician,* **48**, 324-335.

Krewski, D., Leroux, B.G., Bleuer, S.R., and Broekhoven, L.H. (1993). Modeling the Ames *salmonella/*microsome Assay. *Biometrics,* **49**, 499-510.

Krige, D.G. (1951). A statistical approach to *some* basic mine valuation problems on the Witwatersrand. *J. Chemical, Metallurgical and Mining Society of South Africa,* **52**, 119-139.

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.*, **73**, 805-811.

Laird, N.M., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *J. Amer. Statist. Assoc.,* **82**, 97-105.

Laird, N.M. and Louis, T.A. (1982). Approximate posterior distributions for incomplete data problems. *J. Roy. Statist. Soc., Ser. B,* **44**, 190-200.

Laird, N.M. and Louis, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J. Amer. Statist. Assoc.*, **82**, 739-757.

Laird, N.M. and Louis, T.A. (1989). Empirical Bayes confidence intervals for a series of related experiments. *Biometrics,* **45**, 481-495.

Laird, N.M. and Louis, T.A. (1989). Empirical Bayes ranking methods. *J. Educational Statistics*, **14**, 29-46.

Laird, N.M. and Louis, T.A. (1991). Smoothing the non-parameterc estimate of a prior distribution by roughening: A computational study. *Computational Statistics and Data Analysis,* **12**, 27-37.

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika,* **70**, 659-663.

Lange, N., Carlin, B.P., and Gelfand, A.E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion). J. Amer. Statist. Assoc., 87, 615-632.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *Annals of Statistics*, **17**, 624-642.

Laplace, P.S. (1986). Memoir on the probability of the causes of events (English translation of the 1774 French original by S.M. Stigler). *Statistical Science*, **1**, 364-378.

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *J. Roy. Statist. Soc., Ser. B*, **57**, 247-262.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **50**, 157-224.

Lavine, J. and Schervish, M.J. (1999). Bayes factors: what they are and what they are not. *The American Statistician*, **53**, 119-123.

Le, N.D., Sun, W., and Zidek, J.V. (1997). Bayesian multivariate spatial interpolation with data missing by design. *J. Roy. Statist. Soc., Ser. B*, **59**, 501-510.

Lee, P.M. (1997). Bayesian Statistics: An Introduction, 2nd ed. London: Arnold.

Leonard, T. and Hsu, J.S.J. (1999). *Bayesian Methods.* Cambridge: Cambridge University Press.

Liang, K.-Y. and Waclawiw, M.A. (1990). Extension of the Stein estimating procedure through the use of estimating functions. *J. Amer. Statist. Assoc.*, **85**, 435-440.

Lin, D.Y., Fischl, M.A., and Schoenfeld, D.A. (1993). Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. *Statist. in Med.*, **12**, 835-842.

Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.*, **84**, 1074-1078.

Lindley, D.V. (1972). *Bayesian statistics: A review*. Philadelphia: SIAM.

Lindley, D.V. (1983). Discussion of "Parametric empirical Bayes inference: Theory and Applications," by C.N. Morris. *J. Amer. Statist. Assoc.*, **78**, 61-62.

Lindley, D.V. (1990). The 1988 Wald memorial lectures: The present position in Bayesian statistics (with discussion). *Statistical Science*, **5**, 44-89.

Lindley, D.V. and Phillips, L.D. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.*, **30**, 112-119.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc., Ser. B*, **34**, 1-41.

Lindsay, B.G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, **11**, 86-94.

Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673-687.

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika*, **80**, 295-304.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Liu, C.H., Rubin, D.B., and Wu, Y.N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**, 755-770.

Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical*

Association, **89**, 958-966.

Liu, J.S., Wong, W.H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika,* 81, 27-40.

Liu, J.S. and Wu, Y.N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.*, **94**, 1264-1274.

Longmate, J.A. (1990). Stochastic approximation in bootstrap and Bayesian approaches to interval estimation in hierarchical models. In Computing Science and Statistics: *Proc. 22nd Symposium on the Interface*, New York: Springer-Verlag, pp. 446-450.

Louis, T.A. (1981a). Confidence intervals for a binomial parameter after observing no successes. *The American Statistician*, **35**, 151-154.

Louis, T.A. (1981b). Nonparametric analysis of an accelerated failure time model. *Biometrika,* **68**, 381-390.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc., Ser. B*, **44**, 226-233.

Louis, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.*, **79**, 393-398.

Louis, T.A. (1991). Using empirical Bayes methods in biopharmaceutical research (with discussion). *Statistics in Medicine,* **10**, 811-829.

Louis, T.A. and Bailey, J.K. (1990). Controlling error rates using prior information and marginal totals to select tumor sites. *J. Statistical Planning and Inference*, **24**, 297-316.

Louis, T.A. and DerSimonian, R. (1982). Health statistics based on discrete population groups. In *Regional Variations in Hospital Use*, D. Rothberg, ed., Boston: D.C. Heath & Co.

Louis, T.A. and Shen, W. (1999). Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks. *Statistics in Medicine,* **18**, 2493-2505.

MacEachern, S.N. and Berliner, L.M. (1994). Subsampling the Gibbs sampler. *The American Statistician,* **48**, 188-190.

MacEachern, S.N. and Müller, P. (1994). Estimating mixture of Dirichlet process models. Discussion Paper 94-11, Institute for Statistics and Decision Sciences, Duke University.

MacEachern, S.N. and Peruggia, M. (2000a). Subsampling the Gibbs sampler: variance reduction. *Statisitcs and Probability Letters*, **47**, 91-98.

MacEachern, S.N. and Peruggia, M. (2000b). Importance link function estimation for Markov chain Monte Carlo methods. To appear, *Journal of Computational and Graphical Statistics*.

MacGibbon, B. and Tomberlin, T.J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology,* **15**, 237-252.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.*, **89**, 1535-1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215-232.

Manton, K.E., Woodbury, M.A., Stallard, E., Riggan, W.B., Creason, J.P., and

Pellom, A.C. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. *J. Amer. Statist. Assoc.*, **84**, 637-650.

Mann, P., Louis, T.A., Lane, T.J., Gottlieb, L., Engel, P., and Rippey, R.M. (1988). Unfavorable outcomes of drug therapy: Subjective probability versus confidence intervals. *J. Clin. Pharm. and Therapeutics*, **13**, 213-217.

Maritz, J.S. and Lwin, T. (1989). *Empirical Bayes Methods,* 2nd ed. London: Chapman and Hall.

Marriott, J., Ravishanker, N., Gelfand, A., and Pai, J. (1996). Bayesian analysis of ARMA processes: Complete sampling-based inference under exact likelihoods. In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. Berry, K. Chaloner, and J. Geweke, New York: Wiley, pp. 243-256.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models,* 2nd ed. London: Chapman and Hall.

McNeil, A.J. and Gore, S.M. (1996). Statistical analysis of zidovudine (AZT) effect on CD4 cell counts in HIV disease. *Statistics in Medicine,* **15**, 75-92.

Meeden, G. (1972). Some admissible empirical Bayes procedures. *Ann. Math. Statist.*, **43**, 96-101.

Meilijson, I. (1989). A fast improvement of the EM algorithm on its own terms. *J. Roy. Statist. Soc., Ser B,* **51**, 127-138.

Meng, X.-L. (1994). Posterior predictive *p*-values. *Ann. Statist.,* **22**, 1142-1160.

Meng, X.-L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.*, **86**, 899-909.

Meng, X.-L. and Rubin, D.B. (1992). Recent extensions to the EM algorithm (with discussion). In *Bayesian Statistics 4,* J. M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 307- 315.

Meng, X.-L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika,* **80**, 267-278.

Meng, X.-L. and Van Dyk, D. (1997). The EM algorithm - an old folk-song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc., Ser. B,* **59**, 511-567.

Meng, X.-L. and Van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika,* 86, 301-320.

Mengersen, K.L., Robert, C.P., and Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: a reviewww (with discussion). In *Bayesian Statistics 6,* J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 415-440.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics,* **21**, 1087-1091.

Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. London: Springer-Verlag.

Mira, A. (1998). *Ordering, Slicing and Splitting Monte Carlo Markov Chains.* Unpublished Ph.D. thesis, Department of Statistics, University of Minnesota.

Mira, A., Moller, J., and Roberts, G.O. (1999). Perfect slice samplers. Technical report, Department of Statistics, Aalborg University, Denmark.

Mira, A. and Sargent, D. (2000). Strategies for speeding Markov chain Monte

Carlo algorithms. Technical report, Faculty of Economics, University of Insubria, Varese, Italy.

Mira, A. and Tierney, L. (1997). On the use of auxiliary variables in Markov chain Monte Carlo sampling. Technical report, School of Statistics, University of Minnesota.

Mood, A.M., Graybill, F.A., and Boes, D.C. (1974). *Introduction to the Theory of Statistics,* 3rd ed., New York: McGraw-Hill.

Morris, C.N. (1983a). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.*, **78**, 47-65.

Morris, C.N. (1983b). Natural exponential families with quadratic variance functions: Statistical theory. *Ann. Statist.*, **11**, 515-529.

Morris, C.N. (1988). Determining the accuracy of Bayesian empirical Bayes estimators in the familiar exponential families. In *Statistical Decision Theory and Related Topics*, eds. S.S. Gupta and J.O. Berger, New York: Springer-Verlag, 327-344.

Morris, C.N. and Normand, S.L. (1992). Hierarchical models for combining information and for meta-analyses (with discussion). In *Bayesian Statistics 4,* J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 321-344.

Mosteller, F. and Wallace, D.L. (1964). *Inference and Disputed Authorship: The Federalist.* Reading, MA: Addison-Wesley.

Mukhopadhyay, S. and Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. *J. Amer. Statist. Assoc.*, **92**, 633-639.

Mugglin, A.S., Carlin, B.P., and Gelfand, A.E. (2000). Fully model based approaches for spatially misaligned data. To appear, *J. Amer. Statist. Assoc.*

Muller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical Report 91-09, Department of Statistics, Purdue University.

Muller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *J. Amer. Statist. Assoc.*, **90**, 1322-1330.

Myers, L.E., Adams, N.H., Hughes, T.J., Williams, L.R., and Claxton, L.D. (1987). An interlaboratory study of an EPA/ Ames/salmonella test protocol. *Mutation Research,* **182**, 121-133.

Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.*, **90**, 233-241.

Natarajan, R. and McCulloch, C.E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika,* **82**, 639-643.

Natarajan, R. and Kass, R.E. (2000). Reference Bayesian methods for generalized linear mixed models. *J. Amer. Statist. Assoc.*, **95**, 227-237.

National Research Council (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond.* Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, C.F. Citro and G. Kalton, eds. Washington, D.C.: National Academy Press.

Naylor, J.C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214-235.

Neal, R.M. (1996a). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing,* **6**, 353-366.

Neal, R.M. (1996b). *Bayesian Learning for Neural Networks,* New York: Springer-Verlag.

Neal, R.M. (1997). Markov chain Monte Carlo methods based on `slicing' the density function. Technical Report No. 9722, Dept. of Statistics, University of Toronto.

Neal, R.M. (1998). Suppressing random walks in Markov Chain Monte Carlo using ordered overrelaxation. In *Learning in Graphical Models,* M.I. Jordan, ed., Dordrecht: Kluwer Academic Publishers, pp. 205-228.

Newton, M.A., Guttorp, P., and Abkowitz, J.L. (1992). Bayesian inference by simulation in a stochastic model from hematology. *Computing Science and Statistics,* **24**, 449-455.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Roy. Statist. Soc., Ser. B,* **56**, 1-48.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge: Cambridge University Press.

Odell, P.L. and Feiveson, A.H. (1966). A numerical procedure to generate a sample covariance matrix. *J. Amer. Statist. Assoc.,* **61**, 198-203.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics Volume 2b: Bayesian Inference*. London: Edward Arnold.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc., Ser. B,* **57**, 99-138.

O'Hagan, A. and Berger, J.O. (1988). Ranges of posterior probabilities for quasi-unimodal priors with specified quantiles. *J. Am. Statist. Assoc.,* **83**, 503-508.

Pauler, D.K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika,* **85**, 13-27.

Phillips, D.B. and Smith, A.F.M. (1996). Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice,* eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, London: Chapman and Hall, pp. 215-239.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

Pole, A., West, M., and Harrison, P.J. (1994). *Applied Bayesian Forecasting and Time Series Analysis.* London: Chapman and Hall.

Polson, N.G. (1996). Convergence of Markov chain Monte Carlo algorithms (with discussion). In *Bayesian Statistics 5,* eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, Oxford: Oxford University Press, pp. 297-322.

Prentice, R.L. (1976). A generalization of the probit and logit model for dose response curves. *Biometrics,* **32**, 761-768.

Press, S.J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference,* 2nd ed. New York: Krieger.

Propp, J.G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms,* **9**, 223-252.

Racine, A., Grieve, A.P., Fluher, H., and Smith, A.F.M. (1986). Bayesian methods in practice: Experiences in the pharmaceutical industry. *Applied Statistics,* **35**, 93-120.

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251-266.

Raftery, A.E. and Lewis, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 763-773.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, **92**, 179-191.

Richardson, S. and Green, P.J. (1997). Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc., Ser. B*, **59**, 731-758

Ripley, B.D. (1987). Stochastic Simulation. New York: Wiley.

Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy Gibbs sampler. *J. Amer. Statist. Assoc.*, **87**, 861-868.

Robbins, H. (1955). An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Symp. on Math. Statist. and Prob.*, **1**, Berkeley, CA: Univ. of California Press, pp. 157-164.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.*, **1**,713-723.

Robert, C.P. (1994). *The Bayesian Choice: A Decision-Theoretic Motivation.* New York: Springer-Verlag.

Robert, C.P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, **10**, 231-253.

Robert, C.P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* New York: Springer-Verlag.

Roberts, G.O. and Rosenthal, J.S. (1999a). Convergence of slice samper Markov chains. *J. Roy. Statist. Soc., Ser. B*, **61**, 643-660.

Roberts, G.O. and Rosenthal, J.S. (1999b). The polar slice sampler. Technical report, Department of Mathematics and Statistics, Lancaster University.

Roberts, G.O. and Sahu, S.K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc., Ser. B*, **59**, 291-317.

Roberts, G.O. and Smith, A.F.M. (1993). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, **49**, 207-216.

Roberts, G.O. and Tweedie, R.L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83, 95-110.

Robertson, T., Wright, F.T., and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Chichester: John Wiley & Sons.

Rosenthal, J.S. (1993). Rates of convergence for data augmentation on finite sample spaces. *Ann. App. Prob.*, **3**, 819-839.

Rosenthal, J.S. (1995a). Rates of convergence for Gibbs sampling for variance component models. *Ann. Statist.*, **23**, 740-761.

Rosenthal, J.S. (1995b). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, **90**, 558-566.

Rosenthal, J.S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing*, **6**, 269-275.

Rubin, D.B. (1980). Using empirical Bayes techniques in the law school validity studies (with discussion). *J. Amer. Statist. Assoc.*, **75**, 801-827.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151-1172.

Rubin, D.B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussion). In *Bayesian* Statistics 3, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 395-402.

Samaniego, F.J. and Reneau, D.M. (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *J. Amer. Statist. Assoc.*, **89**, 947-957.

Sargent, D.J. (1995). A general framework for random effects survival analysis in the Cox proportional hazards setting. Research Report 95-004, Division of Biostatistics, University of Minnesota.

Sargent, D.J. and Carlin, B.P. (1996). Robust Bayesian design and analysis of clinical trials via prior partitioning (with discussion). In *Bayesian Robustness,* IMS Lecture Notes - Monograph Series, **29**, eds. J.O. Berger et al., Hayward, CA: Institute of Mathematical Statistics, pp. 175-193.

Sargent, D.J., Hodges, J.S., and Carlin, B.P. (2000). Structured Markov chain Monte Carlo. To appear, *Journal of Computational and Graphical Statistics*.

Schervish, M.J. and Carlin, B.P. (1992). On the convergence of successive substitution sampling. *J. Computational and Graphical Statistics*, **1**, 111-127.

Schluchter, M.D. (1988). "5V: Unbalanced repeated measures models with structured covariance matrices," in *BMDP Statistical Software Manual, Vol. 2,* W.J. Dixon, ed., pp. 1081-1114.

Schmeiser, B. and Chen, M.H. (1991). On random-direction Monte Carlo sampling for evaluating multidimensional integrals. Technical Report #91-39, Department of Statistics, Purdue University.

Schmittlein, D.C. (1989). Surprising inferences from unsurprising observations: Do conditional expectations really regress to the mean? *The American Statistician,* **43**, 176-183.

Schumacher, M., Olschewski, M., and Schmoor, C. (1987). The impact of heterogeneity on the comparison of survival times. Statist. in *Med.,* 6, 773-784.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*. New York: John Wiley & Sons.

Shen, W. and Louis, T.A. (1998). Triple-goal estimates in two-stage, hierarchical models. *J. Roy. Statist. Soc., Ser. B*, **60**, 455-471.

Shen, W. and Louis, T.A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *J. Comp. Graph. Statist.*, **8**, 800-823.

Shen, W. and Louis, T.A. (2000). Triple-goal estimates for disease mapping. To appear, Statistics in *Medicine*.

Shih, J. (1995). Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clinical Trials*, **16**, 395-407.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.*, **4**, 1200-1209.

Sinha, D. (1993). Semiparametric Bayesian analysis of multiple event time data. *J. Amer. Statist. Assoc.*, **88**, 979-983.

Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, **46**, 84-88.

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc., Ser. B*, **55**, 1-24.

Snedecor, G.W. and Cochran, W.G. (1980). *Statistical Methods,* seventh edition. Ames, Iowa: The Iowa State University Press.

Spiegelhalter, D.J. (2000). Bayesian methods for cluster randomised trials with continuous responses. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Spiegelhalter, D.J., Best, N., and Carlin, B.P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009, Division of Biostatistics, University of Minnesota.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., and Cowell, R.G. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science,* **8**, 219-283.

Spiegelhalter, D.J., Freedman, L.S., and Parmar, M.K.B. (1994). Bayesian approaches to randomised trials (with discussion). *J. Roy. Statist. Soc., Ser. A*, **157**, 357-416.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R., and Abrams, K.R. (2000). Bayesian methods in health technology assessment. To appear, *Health Technology Assessment.*

Spiegelhalter, D.J., Thomas, A., Best, N., and Gilks, W.R. (1995a). BUGS: Bayesian inference using Gibbs sampling, Version 0.50. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Spiegelhalter, D.J., Thomas, A., Best, N., and Gilks, W.R. (1995b). BUGS examples, Version 0.50. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. on Math. Statist. and Prob.*, **1**, Berkeley, CA: Univ. of California Press, pp. 197-206.

Stein, C. (1981). Estimation of the parameters of a multivariate normal distribution: I. Estimation of the means. *Annals of Statistics,* **9**, 1135-1151.

Stephens, D.A. and Smith, A.F.M. (1992). Sampling-resampling techniques for the computation of posterior densities in normal means problems. *Test*, **1**, 1-18.

Stern, H.S. and Cressie, N. (1999). Inference for extremes in disease mapping. In *Disease Mapping and Risk Assessment for Public Health*, A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, and R. Bertollini, eds., New York: Wiley, pp. 63-84.

Stoffer, D.S., Scher, M.S., Richardson, G.A., Day, N.L., and Coble, P.A. (1988). A Walsh-Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *J. Amer. Statist. Assoc.*, **83**, 954-963.

Swendsen, R.H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte

Carlo simulations. Phys. *Rev. Letters,* **58**, 86-88.

Tamura, R.N. and Young, S.S. (1986). The incorporation of historical control information in tests of proportions: Simulation study of Tarone's procedure. *Biometrics,* **42**, 343-349.

Tanner, M.A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd ed. New York: Springer-Verlag.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, **82**, 528-550.

Tarone, R.E. (1982). The use of historical control information in testing for a trend in proportions. *Biometrics*, **38**, 215-220.

Ten Have, T.R. and Localio, A.R. (1999). Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics*, **55**, 1022-1029.

Thisted, R.A. (1988). Elements of Statistical Computing. London: Chapman and Hall.

Tierney, L. (1990). LISP-STAT - An Object-Oriented Environment for Statistical Computing and Dynamic Graphics. New York: John Wiley & Sons.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, **22**, 1701-1762.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82-86.

Tierney, L., Kass, R.E., and Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.*, **84**, 710-716.

Titterington, D.M., Makov, U.E. and Smith, A.F.M. (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester: John Wiley & Sons.

Tolbert, P., Mulholland, J., Macintosh, D., Xu, F., Daniels, D., Devine, O., Carlin, B.P., Butler, A., Nordenberg, D., and White, M. (2000). Air quality and pediatric emergency room visits for asthma in Atlanta. *Amer. J. Epidemiology,* **151:8**, 798-810.

Tomberlin, T.J. (1988). Predicting accident frequencies for drivers classified by two factors. *J. Amer. Statist. Assoc.*, **83**, 309-321.

Treloar, M.A. (1974). Effects of Puromycin on galactosyltransferase of golgi membranes. Unpublished Master's thesis, University of Toronto.

Troughton, P.T. and Godsill, S.J. (1997). A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. Technical Report CUED/F-INFENG/TR.304, Engineering Department, Cambridge University.

Tsiatis, A.A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*, **18**, 354-372.

Tsutakawa, R.K., Shoop, G.L., and Marienfeld, C.J. (1985). Empirical Bayes estimation of cancer mortality rates. *Statist. in Med.*, **4**, 201-212.

Tukey, J.W. (1974). Named and faceless values: An initial exploration in memory of Prasanta C. Mahalanobis. *Sankhya, Series A*, **36**, 125-176.

van Houwelingen, J.C. (1977). Monotonizing empirical Bayes estimators for a class of discrete distributions with monotone likelihood ratio. *Statist. Neerl.*,

31, 95-104.

van Houwelingen, J.C. and Thorogood, J. (1995). Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine,* **14**, 1999-2008.

van Ryzin, J. and Susarla, J.V. (1977). On the empirical Bayes approach to multiple decision problems. *Ann. Statist*., **5**, 172-181.

Vaupel, J.W., Manton, K.G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography,* **16**, 439-454.

Vaupel, J.W. and Yashin, A.I. (1985). Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *The American Statistician,* **39**, 176-185.

Verdinelli, I. (1992). Advances in Bayesian experimental design (with discussion). In *Bayesian Statistics 4,* J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 467-481.

Von Mises, R. (1942). On the correct use of Bayes' formula. *Ann. Math. Statist.*, **13**, 156-165.

Wagener, D.K. and Williams, D.R. (1993). Equity in environmental health: Data collection and interpretation issues. *Toxicology and Industrial Health,* **9**, 775-795.

Wakefield, J. (1996). The Bayesian analysis of population pharmacokinetic models. *J. Amer. Statist. Assoc.*, **91**, 62-75.

Wakefield, J.C. (1998). Discussion of "Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics," by J.S. Hodges. *J. Roy. Stat. Soc., Ser. B*, 60, 523-525.

Wakefield, J.C., Gelfand, A.E., and Smith, A.F.M. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing,* **1**, 129-133.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A., and Gelfand, A.E. (1994). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *J. Roy. Statist. Soc., Ser. C (Applied Statistics)*, **43**, 201-221.

Walker, S. (1995). Generating random variates from D-distributions via substitution sampling. *Statistics in Computing*, **5**, 311-315.

Waller, L.A., Carlin, B.P., Xia, H., and Gelfand, A.E. (1997). Hierarchical spatio-temporal mapping of disease rates. *J. Amer. Statist. Assoc*., **92**, 607-617.

Waller, L.A., Turnbull, B.W., Clark, L.C., and Nasca, P. (1994). Spatial pattern analyses to detect rare disease clusters. In *Case Studies in Biometry*, N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse, eds., New York: John Wiley & Sons, pp. 3-23.

Walter, S.D. and Birnie, S.E. (1991). Mapping mortality and morbidity patterns: An international comparison. *International J. Epidemiology*, **20**, 678-689.

Ware, J.H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion). *Statistical Science*, **4**, 298-340.

Wasserman, L. and Kadane, J.B. (1992). Computing bounds on expectations. *J. Amer. Statist. Assoc*., **87**, 516-522.

Waternaux, C., Laird, N.M., and Ware, J.A. (1989). Methods for analysis of longitudinal data: Blood lead concentrations and cognitive development. *J. Amer. Statist. Assoc*., **84**, 33-41.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.*, **85**, 699-704.

West, M. (1992). Modelling with mixtures (with discussion). In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 503-524.

West, M., Müller, P., and Escobar, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D. V. Lindley,* eds. A.F.M. Smith and P.R. Freeman, London: Wiley.

West, M. and Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag.

West, M., Harrison, P.J., and Pole, A. (1987). BATS: Bayesian Analysis of Time Series. *The Professional Statistician*, **6**, 43-46.

Williams, E. (1959). *Regression Analysis*. New York: Wiley.

Whittemore, A.S. (1989). Errors-in-variables regression using Stein estimates. *The American Statistician*, **43**, 226-228.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika,* **80**, 791-795.

Wolfinger, R. and Kass, R.E. (2000). Non-conjugate Bayesian analysis of variance component models. To appear, *Biometrics*.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233-243.

Wolfinger, R.D. and Rosner, G.L. (1996). Bayesian and frequentist analyses of an in vivo experiment in tumor hemodynamics. In *Bayesian Biostatistics,* eds. D. Berry and D. Stangl, New York: Marcel Dekker, pp. 389-410.

Wolfson, L.J. (1995). Elicitation of priors and utilities for Bayesian analysis. Unpublished Ph.D. disssertation, Department of Statistics, Carnegie Mellon University.

Wolpert, R.L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika,* **85**, 251-267.

Wooff, D. (1992). [B/D] works. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 851-859.

Xia, H., Carlin, B.P., and Waller, L.A. (1997). Hierarchical models for mapping Ohio lung cancer rates. *Environmetrics*, **8**, 107-120.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120-131.

Zaslavsky, A.M. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *J. Amer. Statist. Assoc.*, **88**, 1092-1105.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.*, **86**, 79-86.

Zhou, X.-H., Perkins, A.J., and Hui, S.L. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician*, **53**, 282-290.