# Preface

This book was written slowly over the course of the last five years. During that time, a number of advances have been made in Bayesian statistics and Markov chain Monte Carlo (MCMC) methods, but, in my opinion, the market still lacks a *truly* introductory book written explicitly for social scientists that thoroughly describes the actual process of Bayesian analysis using these methods. To be sure, a variety of introductory books are available that cover the basics of the Bayesian approach to statistics (e.g., Gill 2002 and Gelman et al. 1995) and several that cover the foundation of MCMC methods (e.g., beginning with Gilks et al. 1996). Yet, a highly applied book showing how to use MCMC methods to complete a Bayesian analysis involving typical social science models applied to typical social science data is still sorely lacking. The goal of this book is to fill this niche.

The Bayesian approach to statistics has a long history in the discipline of statistics, but prior to the 1990s, it held a marginal, almost cult-like status in the discipline and was almost unheard of in social science methodology. The primary reasons for the marginal status of the Bayesian approach include (1) philosophical opposition to the use of "prior distributions" in particular and the subjective approach to probability in general, and (2) the lack of computing power for completing realistic Bayesian analyses. In the 1990s, several events occurred simultaneously to overcome these concerns. First, the explosion in computing power nullified the second limitation of conducting Bayesian analyses, especially with the development of sampling based methods (e.g., MCMC methods) for estimating parameters of Bayesian models. Second, the growth in availability of longitudinal (panel) data and the rise in the use of hierarchical modeling made the Bayesian approach more appealing, because Bayesian statistics offers a natural approach to constructing hierarchical models. Third, there has been a growing recognition both that the enterprise of statistics is a subjective process in general and that the use of prior distributions need not influence results substantially. Additionally, in many problems, the use of a prior distribution turns out to be advantageous.

The publication of Gelfand and Smith's 1990 paper describing the use of MCMC simulation methods for summarizing Bayesian posterior distributions was the watershed event that launched MCMC methods into popularity in statistics. Following relatively closely on the heels of this article, Gelman et al.'s (1995) book, *Bayesian Data Analysis*, and Gilks et al.'s (1996) book, *Markov Chain Monte Carlo in Practice*, placed the Bayesian approach in general, and the application of MCMC methods to Bayesian statistical models, squarely in the mainstream of statistics. I consider these books to be classics in the field and rely heavily on them throughout this book.

Since the mid-1990s, there has been an explosion in advances in Bayesian statistics and especially MCMC methodology. Many improvements in the recent past have been in terms of (1) monitoring and improving the performance of MCMC algorithms and (2) the development of more refined and complex Bayesian models and MCMC algorithms tailored to specific problems. These advances have largely escaped mainstream social science.

In my view, these advances have gone largely unnoticed in social science, because purported introductory books on Bayesian statistics and MCMC methods are not truly introductory for this audience. First, the mathematics in introductory books is often too advanced for a mainstream social science audience, which begs the question: "introductory *for whom?*" Many social scientists do not have the probability theory and mathematical statistics background to follow many of these books beyond the first chapter. This is not to say that the material is impossible to follow, only that more detail may be needed to make the text and examples more readable for a mainstream social science audience.

Second, many examples in introductory-level Bayesian books are at best foreign and at worst irrelevant to social scientists. The probability distributions that are used in many examples are not typical probability distributions used by social scientists (e.g., Cauchy), and the data sets that are used in examples are often atypical of social science data. Specifically, many books use small data sets with a limited number of covariates, and many of the models are not typical of the regression-based approaches used in social science research. This fact may not seem problematic until, for example, one is faced with a research question requiring a multivariate regression model for 10,000 observations measured on 5 outcomes with 10 or more covariates. Nonetheless, research questions involving large-scale data sets are not uncommon in social science research, and methods shown that handle a sample of size 100 measured on one or two outcomes with a couple of covariates simply may not be directly transferrable to a larger data set context. In such cases, the analyst without a solid understanding of the linkage between the model and the estimation routine may be unable to complete the analysis. Thus, some discussion tailored to the practicalities of *real* social science data and computing is warranted.

Third, there seems to be a disjunction between introductory books on Bayesian theory and introductory books on applied Bayesian statistics. One

of the greatest frustrations for me, while I was learning the basics of Bayesian statistics and MCMC estimation methods, was (and is) the lack of a book that links the theoretical aspects of Bayesian statistics and model development with the application of modern estimation methods. Some examples in extant books may be substantively interesting, but they are often incomplete in the sense that discussion is truncated after model development without adequate guidance regarding how to estimate parameters. Often, suggestions are made concerning how to go about implementing only certain aspects of an estimation routine, but for a person with no experience doing this, these suggestions are not enough.

In an attempt to remedy these issues, this book takes a step back from the most recent advances in Bayesian statistics and MCMC methods and tries to bridge the gap between Bayesian theory and modern Bayesian estimation methods, as well as to bridge the gap between Bayesian statistics books written as "introductory" texts for statisticians and the needs of a mainstream social science audience. To accomplish this goal, this book presents very little that is new. Indeed, most of the material in this book is now "old-hat" in statistics, and many references are a decade old (In fact, a second edition of Gelman et al.'s 1995 book is now available). However, the trade-off for not presenting much new material is that this book explains the process of Bayesian statistics and modern parameter estimation via MCMC simulation methods in great depth. Throughout the book, I painstakingly show the modeling process from model development, through development of an MCMC algorithm to estimate its parameters, through model evaluation, and through summarization and inference.

Although many introductory books begin with the assumption that the reader has a solid grasp of probability theory and mathematical statistics, I do not make that assumption. Instead, this book begins with an exposition of the probability theory needed to gain a solid understanding of the statistical analysis of data. In the early chapters, I use contrived examples applied to (sometimes) contrived data so that the forest is not lost for the trees: The goal is to provide an understanding of the issue at hand rather than to get lost in the idiosyncratic features of real data. In the latter chapters, I show a Bayesian approach (or approaches) to estimating some of the most common models in social science research, including the linear regression model, generalized linear models (specifically, dichotomous and ordinal probit models), hierarchical models, and multivariate models.

A consequence of this choice of models is that the parameter estimates obtained via the Bayesian approach are often very consistent with those that could be obtained via a classical approach. This may make a reader ask, "then what's the point?" First, there are many cases in which a Bayesian approach and a classical approach will not coincide, but from my perspective, an introductory text should establish a foundation that can be built upon, rather than beginning in unfamiliar territory. Second, there are additional benefits to taking a Bayesian approach beyond the simple estimation of model

parameters. Specificially, the Bayesian approach allows for greater flexibility in evaluating model fit, comparing models, producing samples of parameters that are not directly estimated within a model, handling missing data, "tweaking" a model in ways that cannot be done using canned routines in existing software (e.g., freeing or imposing constraints), and making predictions/forecasts that capture greater uncertainty than classical methods. I discuss each of these benefits in the examples throughout the latter chapters.

Throughout the book I thoroughly flesh out each example, beginning with the development of the model and continuing through to developing an MCMC algorithm (generally in R) to estimate it, estimating it using the algorithm, and presenting and summarizing the results. These programs should be straightforward, albeit perhaps tedious, to replicate, but some programming is inherently required to conduct Bayesian analyses. However, once such programming skills are learned, they are incredibly freeing to the researcher and thus well worth the investment to acquire them. Ultimately, the point is that the examples are thoroughly detailed; nothing is left to the imagination or to guesswork, including the mathematical contortions of simplifying posterior distributions to make them recognizable as known distributions.

A key feature of Bayesian statistics, and a point of contention for opponents, is the use of a prior distribution. Indeed, one of the most complex things about Bayesian statistics is the development of a model that includes a prior and yields a "proper" posterior distribution. In this book, I do not concentrate much effort on developing priors. Often, I use uniform priors on most parameters in a model, or I use "reference" priors. Both types of priors generally have the effect of producing results roughly comparable with those obtained via maximum likelihood estimation (although not in interpretation!). My goal is not to minimize the importance of choosing appropriate priors, but instead it is not to overcomplicate an introductory exposition of Bayesian statistics and model estimation. The fact is that most Bayesian analyses explicitly attempt to minimize the effect of the prior. Most published applications to date have involved using uniform, reference, or otherwise "noninformative" priors in an effort to avoid the "subjectivity" criticism that historically has been levied against Bayesians by classical statisticians. Thus, in most Bayesian social science research, the prior has faded in its importance in differentiating the classical and Bayesian paradigms. This is not to say that prior distributions are unimportant—for some problems they may be very important or useful—but it is to say that it is not necessary to dwell on them.

The book consists of a total of 11 chapters plus two appendices covering (1) calculus and matrix algebra and (2) the basic concepts of the Central Limit Theorem. The book is suited for a highly applied one-semester graduate level social science course. Each chapter, including the appendix but excluding the introduction, contains a handful of exercises at the end that test the understanding of the material in the chapter at both theoretical and applied levels. In the exercises, I have traded quantity for quality: There are relatively few exercises, but each one was chosen to address the essential material in

the chapter. The first half of the book (Chapters 1-6) is primarily theoretical and provides a generic introduction to the theory and methods of Bayesian statistics. These methods are then applied to common social science models and data in the latter half of the book (Chapters 7-11). Chapters 2-4 can each be covered in a week of classes, and much of this material, especially in Chapters 2 and 3, should be review material for most students. Chapters 5 and 6 will most likely each require more than a week to cover, as they form the nuts and bolts of MCMC methods and evaluation. Subsequent chapters should each take 1-2 weeks of class time. The models themselves should be familiar, but the estimation of them via MCMC methods will not be and may be difficult for students without some programming and applied data analysis experience. The programming language used throughout the book is R, a freely available and common package used in applied statistics, but I introduce the program WinBugs in the chapter on hierarchical modeling. Overall, R and WinBugs are syntactically similar, and so the introduction of WinBugs is not problematic. From my perspective, the main benefit of WinBugs is that some derivations of conditional distributions that would need to be done in order to write an R program are handled automatically by WinBugs. This feature is especially useful in hierarchical models. All programs used in this book, as well as most data, and hints and/or solutions to the exercises can be found on my Princeton University website at: www.princeton.edu/∼slynch.

## Acknowledgements

I have a number of people to thank for their help during the writing of this book. First, I want to thank German Rodriguez and Bruce Western (both at Princeton) for sharing their advice, guidance, and statistical knowledge with me as I worked through several sections of the book. Second, I thank my friend and colleague J. Scott Brown for reading through virtually all chapters and providing much-needed feedback over the course of the last several years. Along these same lines, I thank Chris Wildeman and Steven Shafer for reading through a number of chapters and suggesting ways to improve examples and the general presentation of material. Third, I thank my statistics thesis advisor, Valen Johnson, and my mentor and friend, Ken Bollen, for all that they have taught me about statistics. (They cannot be held responsible for the fact that I may not have learned well, however). For their constant help and tolerance, I thank Wayne Appleton and Bob Jackson, the senior computer folks at Princeton University and Duke University, without whose support this book could not have been possible. For their general support and friendship over a period including, but not limited to, the writing of this book, I thank Linda George, Angie O'Rand, Phil Morgan, Tom Espenshade, Debby Gold, Mark Hayward, Eileen Crimmins, Ken Land, Dan Beirute, Tom Rice, and John Moore. I also thank my son, Tyler, and my wife, Barbara, for listening to me ramble incessantly about statistics and acting as a sounding

board during the writing of the book. Certainly not least, I thank Bill McCabe for helping to identify an egregious error on page 364. Finally, I want to thank my editor at Springer, John Kimmel, for his patience and advice, and I acknowledge support from NICHD grant R03HD050374-01 for much of the work in Chapter 10 on multivariate models.

Despite having all of these sources of guidance and support, all the errors in the book remain my own.

Princeton University *Scott M. Lynch*
April 2007