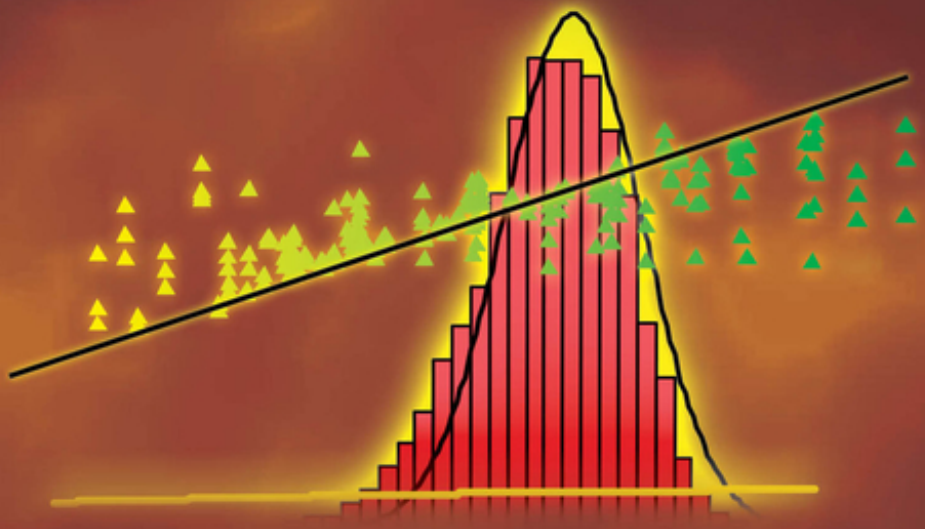
WILEY SERIES IN PROBABILITY AND STATISTICS

Second Edition

Bayesian Statistics and Marketing



Peter Rossi, Greg Allenby
and Sanjog Misra

WILEY

Bayesian Statistics and Marketing

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

Bayesian Statistics and Marketing

Second Edition

Peter E. Rossi
Anderson School of Management
UCLA
Los Angeles
USA

Greg M. Allenby
Ohio State University
Fisher College of Business
Columbus
USA

Sanjog Misra
University of Chicago
Booth School of Business
Chicago
USA

WILEY

This edition first published 2024

Edition History

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data:

Names: Rossi, Peter E. (Peter Eric), 1955- author. | Allenby, Greg M. (Greg Martin), 1956- author. | Misra, Sanjog, author.
Title: Bayesian statistics and marketing / Peter E. Rossi, Greg M. Allenby, Sanjog Misra.
Description: Second edition. | Hoboken, NJ : Wiley, [2024] | Series: Wiley series in probability and statistics | Includes index.
Identifiers: LCCN 2024003452 (print) | LCCN 2024003453 (ebook) | ISBN 9781394219117 (hardback) | ISBN 9781394219131 (adobe pdf) | ISBN 9781394219124 (epub)
Subjects: LCSH: Marketing research–Mathematical models. | Marketing–Mathematical models. | Bayesian statistical decision theory.
Classification: LCC HF5415.2 .R675 2024 (print) | LCC HF5415.2 (ebook) | DDC 658.8/3015118–dc23/eng/20240221
LC record available at https://lccn.loc.gov/2024003452
LC ebook record available at https://lccn.loc.gov/2024003453

Cover Design: Wiley
Cover Image: Courtesy of Peter Rossi, Greg Allenby and Sanjog Misra

Set in 10/12pt Galliard Std by Straive, Chennai, India

To Aimee, Tricia, and Debra

Contents

1
Introduction

Abstract

While the conceptual appeal of Bayesian methods has long been recognized, the recent popularity stems from computational and modeling breakthroughs that have made Bayesian methods attractive for many marketing problems. This book provides a self-contained and comprehensive treatment of Bayesian methods and the marketing problems for which these methods are especially appropriate. It presents a treatment of Bayesian methods that emphasizes the unique aspects of their application to marketing problems. The book emphasizes the unique aspects of the modeling problem in marketing and the modifications of method and models that researchers in marketing have devised. It also provides the requisite methodological knowledge and an appreciation of how these methods can be used to allow the reader to devise and analyze new models. The book takes a stand on customer differences by modeling differences via a probability distribution.

The past 30 years have seen a dramatic increase in the use of Bayesian methods in marketing. Bayesian analyses have been conducted over a wide range of marketing problems from new product introduction to pricing, and with a wide variety of different data sources. While the conceptual appeal of Bayesian methods has long been recognized, the recent popularity stems from computational and modeling breakthroughs that have made Bayesian methods attractive for many marketing problems. This book aims to provide a self-contained and comprehensive treatment of Bayesian methods and the marketing problems for which these methods are especially appropriate. There are unique aspects of important problems in marketing that make particular models and specific Bayesian methods attractive. We, therefore, do not attempt to provide a generic treatment of Bayesian methods. We refer the interested reader to classic treatments by Robert and Casella [2004], Gelman et al. [2004], and Berger [1985] for more general-purpose discussion of Bayesian methods. Instead, we provide a treatment of Bayesian methods that emphasizes the unique aspects of their application to marketing problems.

Bayesian Statistics and Marketing, Second Edition. Peter E. Rossi, Greg M. Allenby, and Sanjog Misra

Until the mid-1980s, Bayesian methods appeared impractical since the class of models for which the posterior inference could be computed was no larger than the class of models for which exact sampling results were available. Moreover, the Bayes approach does require assessment of a prior which some feel to be an extra cost. Simulation methods, in particular Markov Chain Monte Carlo (MCMC) methods, have freed us from computational constraints for a very wide class of models. MCMC methods are ideally suited for models built from a sequence of conditional distributions, often called hierarchical models. Bayesian hierarchical models offer tremendous flexibility and modularity and are particularly useful for marketing problems.

There is an important interaction between the availability of inference methods and the development of statistical models. Nowhere has this been more evident than in the application of hierarchical models to marketing problems. Hierarchical models are those built up through a sequence of conditional distributions. These models match rather closely the various levels at which marketing decisions are made – from individual consumers to the marketplace. Bayesian researchers in marketing have expanded on the standard set of hierarchical models to provide models useful for marketing problems. Throughout this book, we will emphasize the unique aspects of the modeling problem in marketing and the modifications of method and models that researchers in marketing have devised. We hope to provide the requisite methodological knowledge and an appreciation of how these methods can be used to allow the reader to devise and analyze new models. This departs, to some extent, from the standard model of a treatise in statistics in which one writes down a set of models and catalogues the set of methods appropriate for analysis of these models.

1.1 A BASIC PARADIGM FOR MARKETING PROBLEMS

Ultimately, marketing data results from customers taking actions in a particular context and facing a particular environment. The marketing manager can influence some aspects of this environment. Our goal is to provide models of these decision processes and then make optimal decisions conditional on these models. Fundamental to this prospective is that customers are different in their needs and wants for marketplace offerings, thus expanding the set of actions that can be taken. At the extreme, actions can be directed at specific individuals. Even if one-on-one interaction is not possible, the models and system of inference must be flexible enough to admit nonuniform actions.

Once the researcher acknowledges the existence of differences between customers, the modeling task expands to include a model of these differences. Throughout this book, we will take a stand on customer differences by modeling differences via a probability distribution. Those familiar with standard econometric methods will recognize this as related to a random coefficients approach. The primary difference is that we do not regard the customer level parameters as nuisance parameters but, instead, regard these parameters as the goal of inference. Inferences about customer differences are required for any marketing action, from strategic decisions associated with formulating offerings to tactical decisions of customizing prices. Individuals who are most likely to respond to these variables are those that find highest value in the offering's attributes and those that are most price sensitive, neither of whom are well described by parameters such as the mean of the random coefficients distribution.

Statistical modeling of marketing problems consists of three components:

(i) Within-unit behavior

(ii) Across-unit behavior

(iii) Action

"Unit" refers to the particular level of aggregation dictated by the problem and data availability. In many instances, the unit is the consumer. However, it is possible to consider both less and more aggregate levels of analyses. For example, one might consider a particular consumption occasion or survey instances as the "unit" and consider changes in preferences across occasions or over time as part of the model (an example of this is in Yang et al. [2002]). In marketing practice, decisions are often made at a much higher level of aggregation such as the "key account" or sales territory. In all cases, we consider the "unit" as the lowest level of aggregation considered explicitly in the model.

The first component of problem is the conditional likelihood for the "unit-level behavior." We condition on unit-specific parameters that are regarded as the sole source of between-unit differences. The second component is a distribution of these unit-specific parameters over the population of units. Finally, the decision problem is the ultimate goal of modeling exercise. We typically postulate a profit function and ask – what is the optimal action conditional on the model and the information in the data? Given this view of marketing problems, it is natural to consider the Bayesian approach to inference, which provides a unified treatment of all three components.

1.2 A SIMPLE EXAMPLE

As an example of the components outlined in Section 1.1, consider the case of consumers observed making choices between different products. Products are characterized by some vector of choice attribute variables that might include product characteristics, prices, and advertising. Consumers could be observed to make choices either in the marketplace or in a survey/experimental setting. We want to predict how consumers will react to a change in the marketing mix variables or in the product characteristics. Our ultimate goal is to design products or vary the marketing mix so as to optimize profitability.

We start with the "within-unit" model of choice conditional on the observed attributes for each of the choice alternatives. A standard model for this situation is the Multinomial Logit model.

$$\Pr\left[i \middle| x_1, \ldots, x_p, \beta\right] = \frac{\exp\left(x_i'\beta\right)}{\sum_{j=1}^{p} \exp\left(x_j'\beta\right)} \tag{1.2.1}$$

If we observe more than one observation per consumer, it is natural to consider a model that accommodates differences between consumers. That is, we have some information about each consumer's preferences and we can start to tease out these differences. However, we must recognize that in many situations, we have only a small amount of information about each consumer. To allow for the possibility that each consumer has different preferences for attributes, we index the β vectors by c for consumer c.

Given the small amount of information for each consumer, it is impractical to estimate separate and independent logits for each of the C consumers. For this reason, it is useful to think about a distribution of coefficient vectors across the populations of consumers. One simple model would be to assume that the βs are distributed normally over consumers.

$$\beta_c \sim N\left(\mu, V_\beta\right) \tag{1.2.2}$$

One common use of logit models is to compute the implication of changes in marketing actions for aggregate market shares. If we want to evaluate the effect on market share for a change in x for alternative i, then we need to integrate over the distribution in (1.2.1). For a market with a large number of consumers, we might view the expected probability as market share and compute the derivative of market share with respect to an element of x.[1]

$$\frac{\partial MS(i)}{\partial x_{i,j}} = \frac{\partial}{\partial x_{i,j}} \int \Pr\left[i \left| x_1, \ldots, x_p, \beta\right.\right] \varphi\left(\beta \left| \mu, V_\beta\right.\right) d\beta \tag{1.2.3}$$

Here $\varphi()$ is the multivariate normal density.

The derivatives given in (1.2.3) are necessary to evaluate uniform marketing actions such as changing price in a situation in which all consumers face the same price. However, many marketing actions are aimed at a subset of customers or, in some cases, individual customers. In this situation, it is desirable to have a way of estimating not only the common parameters that drive the distribution of βs across consumers but the individual βs as well.

Thus, our objective is to provide a way of inferring about $\{\beta_1, \ldots, \beta_C\}$ as well as μ, V_β. We also want to use our estimates to derive optimal marketing policies. This will mean to maximize expected profits over the range of possible marketing actions.

$$\max_a E\left[\pi\left(a \left| \Omega\right.\right)\right] \tag{1.2.4}$$

Ω represents the information available about the distribution of the outcomes resulting from marketing actions. Clearly, information about both the distribution of choice given the model parameters as well as information about the parameters will be relevant to selecting the optimal action. Our goal, then, is to adopt a system of inference and decision-making that will make it possible to solve (1.2.4). In addition, we will require that there will be practical ways of implementing this system of inference. By practical, we mean computable for problems of the size which practitioners in marketing encounter.

Through this book, we will consider models similar to the simple case considered here and develop these inference and computational tools. We hope to convince the reader that the Bayesian alternative is the right choice.

[1] Some might object to this formulation of the problem as the aggregate market shares are deterministic functions of x. It is a simple matter to add an additional source of randomness to the shares. We are purposely simplifying matters for expositional purposes.

1.3 BENEFITS AND COSTS OF THE BAYESIAN APPROACH

In the beginning of Chapter 2, we outline the basics of the Bayesian approach to inference and decision-making. There are really no other approaches that can provide a unified treatment of inference and decision as well properly account for parameter and model uncertainty. However compelling the logic is behind the Bayesian approach, it has not been universally adopted. The reason for this is that there are nontrivial costs of adopting the Bayesian perspective. We will argue that some of these "costs" have been dramatically reduced and further that some "costs" are not really costs but are actually benefits.

The traditional view is that Bayesian inference provides the benefits of exact sample results, integration of decision-making, "estimation," "testing," and model selection, and a full accounting of uncertainty. Somewhat more controversial is the view that the Bayesian approach delivers the answer to the right question in the sense that Bayesian inference provides answers conditional on the observed data and not based on the distribution of estimators or test statistics over imaginary samples not observed. Balanced against these benefits are three costs: 1. Formulation of a prior; 2. Requirement of a Likelihood function; and 3. Computation of various integrals required in Bayesian paradigm. Development of various simulation-based methods in recent years has drastically lowered the computational costs of the Bayesian approach. In fact, for many of the models considered in this book, non-Bayesian computations would be substantially more difficult or, in some cases, virtually impossible. Lowering of the computational barrier has resulted in a huge increase in the amount of Bayesian applied work.

In spite of increased computational feasibility or, indeed, even computational superiority of the Bayesian approach, some are still reluctant to use Bayesian methods because of the requirement of a prior distribution. From a purely practical point of view, the prior is yet another requirement that the investigator must meet and this imposes a cost to the use of Bayesian approaches. Others are reluctant to utilize prior information based on concerns of scientific "objectivity." Our answer to those with concerns about "objectivity" is twofold. First, to our minds, scientific standards require that replication is possible. Bayesian inference with explicit priors meets this standard. Secondly, marketing is an applied field which means that the investigator is facing a practical problem often in situations with little information and should not neglect sources of information outside of the current data set.

For problems with substantial data information, priors in a fairly broad range will result in much the same a posteriori inferences. However, in any problem in which the data information to "parameters" ratio is low, priors will matter. In models with unit-level parameters, there is often relatively little data information so that it is vital that the system of inference incorporates even small amounts of prior information. Moreover, many problems in marketing explicitly involve multiple information sets so that the distinction between the sample information and prior information is blurred.

High-dimensional parameter spaces arise due to either the large numbers of "units" or the desire to incorporate flexibility in the form of the model specification. Successful solution of problems with high-dimensional parameter spaces requires additional structure. Our view is that prior information is one exceptionally useful way to impose structure on high-dimensional problems. The real barrier is not the philosophical concern over the use of prior information but the assessment of priors in high-dimensional spaces. We need devices for inducing priors on high-dimensional spaces that incorporate the

desired structure with a minimum of effort in assessment. Hierarchical models are one particularly useful method for assessing and constructing priors over parameter spaces of the sort which routinely arise in marketing problems.

Finally, some have argued that any system of likelihood-based inference is problematic due to concerns regarding mis-specification of the likelihood. Tightly parameterized likelihoods can be misspecified, although the Bayesian is not required to believe that there is a "true" model underlying the data. In practice, a Bayesian can experiment with a variety of parametric models as way of guarding against mis-specification. Modern Bayesian computations and modeling methods make the use of a wide variety of models much easier than in the past. Alternatively, more flexible "non" or "semi" parametric models can be used. All nonparametric models are just high-dimensional models to the Bayesian and this simply underscores the need for prior information and Bayesian methods in general. However, there is a school of thought prominent in econometrics that proposes estimators that are consistent for the set of models outside one parametric class (method of moments procedures are the most common of this type). However, in marketing problems, parameter estimates without a probability model are of little use. In order to solve the decision problem, we require the distribution of outcome measures conditional on our actions. This distribution requires not only point estimates of parameters but a specification of their distribution. If we regard the relevant distribution as part of the parameter space, then this statement is equivalent to the need for estimates of all rather than a subset of model parameters.

In a world with full and perfect information, revealed preference should be the ultimate test of the value of a particular approach to inference. The increased adoption of Bayesian methods in marketing shows that the benefits do outweigh the costs for many problems of interest. However, we do feel that the value of Bayesian methods for marketing problem is underappreciated due to lack of information. We also feel that many of the benefits are as yet unrealized since the models and methods are still to be developed. We hope that this book provides a platform for future work on Bayesian methods in marketing.

1.4 AN OVERVIEW OF METHODOLOGICAL MATERIAL AND CASE STUDIES

Chapters 2 and 3 provide a self-contained introduction to the basic principles of Bayesian inference and computation. A background in basic probability and statistics on the level of Casella and Berger [2002] is required to understand this material. We assume a familiarity with matrix notation and basic matrix operations, including the Cholesky root. Those who need a refresher or a concise summary of the relevant material might examine appendices A and B of Koop [2003]. We will develop some of the key ideas regarding joint, conditional, and marginal densities in the beginning of Chapter 2 as we have found that this is an area not emphasized sufficiently in standard mathematical statistics or econometrics courses.

We recognize that a good deal of the material in Chapters 2 and 3 is available in many other scattered sources but we have not found a reference that puts it together in a way that is useful for those interested in marketing problems. We also will include some of the insights that we have obtained from the application of these methods.

Chapters 4 and 5 develop models for within-unit and across-unit analysis. We pay extensive attention to models for discrete data as much disaggregate marketing data involve aspects of discreteness. We also develop the basic hierarchical approach to modeling differences across units and illustrate this approach with a variety of different hierarchical structures and priors.

The problem of model selection and decision theory is developed in Chapter 6. We consider the use of the decision-based metric in valuing information sources and show the importance of loss functions in marketing applications.

Chapter 7 treats the important problem of simultaneity. In models with simultaneity, the distinction between dependent and independent variables is lost as the models are often specified as a system of equations which jointly or simultaneously determine the distribution of a vector of random variables conditional on some set of explanatory or exogenous variables. In marketing applications, the marketing mix variables and sales are joint determined given a set of exogenous demand or cost shifters.

Chapter 8 develops a Bayesian perspective on the Machine Learning literature. Basic concepts of shrinkage and model selection are extremely important in making the highly parameterized models used in the ML literature practical and avoid over-fitting. The main distinction between the ML literature and a classical Bayesian approach is that most ML models are fit with approximate methods that are not fully Bayesian. In addition, the ML literature takes an unabashedly predictive approach rather than an inference approach. These are fundamentally different objectives as is discussed in the chapter.

Chapter 9 takes up the important question of how to conduct inference with text data and develops a number of popular models for analysis of text data. This chapter provides an introduction to Bayesian analysis of text data using the Latent Dirichlet Allocation (LDA) model that summarizes respondent text data by way of topic probabilities, which are unique for each respondent, and word probabilities for each topic that are the same across respondents. Topic probabilities summarize the themes present in textual responses and the word probabilities are used to interpret each topic. We discuss variations of the LDA model that can improve the interpretation of topics and show how the vector of topic probabilities can be used to form integrated models of textual response, choice, and scaled response data. A conjoint dataset is used to illustrate the model. We find that the text data helps clarify the origin of demand.

These core chapters are followed by five case studies from our research agenda. These case studies illustrate the usefulness of the Bayesian approach by tackling important problems that involve extensions or elaborations of the material covered in the first eight chapters. Each of the case studies have been rewritten from their original journal form to use a common notation and emphasize the key points of differentiation for each article. Data and code are available for each of the case studies.

1.5 APPROXIMATE BAYES METHODS AND THIS BOOK

Bayesian methods have seen a number of developments over the last decade or two. Among these developments is the development and use of approximate Bayesian techniques. This sub-field has been fueled by the widespread availability of large datasets as well as the emergence of Machine learning which afford complex modeling of such data. In particular, the need for approximate Bayes methods arises from the computational

challenges associated with exact Bayesian inference. Traditional Bayesian methods (such as MCMC) are known for providing a principled way to represent uncertainty in statistical modeling, but they often require intense computational resources. With the advent of big data and complex models, exact computations at this scale have become intractable in many cases, which has led to the search for more efficient and scalable approaches. In essence, the goal of approximate Bayes methods is to develop techniques that can provide tractable solutions without significantly compromising the quality of the inference.

Several approximate techniques have emerged to address the challenges in exact Bayesian inference. For example, approximation ideas such as Bootstrap approaches which use sampling and averaging to approximate posterior quantities have become widely used. In Chapter 8, we will discuss many of these methods in some detail. A notable exception that we do not discuss in this text is variational inference (VI). The core idea of VI is to turn the problem of Bayesian inference into an optimization problem. This is done by introducing a family of distributions (known as the variational family) and finding the distribution within this family that is closest to the true posterior. VI aims to find a tractable distribution that is closest to the true posterior, providing faster convergence but sometimes at the cost of accuracy. The breadth of ideas in this area and limited applications in marketing have kept us from including this topic in our discussions. In part, we also do not cover VI topics here since there are excellent reviews already available (Blei et al. [2017]).

The field of approximate Bayes methods continues to be an active area of research, and the future directions are promising. There is a growing interest in developing techniques that can balance computational efficiency with accuracy, especially in the context of deep learning and complex hierarchical models. The integration of approximate Bayesian methods with other machine learning paradigms and the development of software packages that make these methods accessible to non-experts are also important trends. Research is also focusing on the theoretical understanding of these methods, providing guarantees on their performance, and exploring their applicability in various scientific and industrial domains. The ongoing collaboration between statisticians, computer scientists, and domain experts ensures that approximate Bayes methods will continue to evolve and play a vital role in statistical modeling and data analysis. Future editions of this text may indeed cover these topics in a lot more depth.

1.6 COMPUTING AND THIS BOOK

It is our belief that no book on practical statistical methods can be credible unless the authors have computed all the methods and models contained therein. For this reason, we have imposed the discipline on ourselves that nothing will be included we haven't computed. It is impossible to assess the practical value of a method without applying it in a realistic setting. Far too often, treatises on statistical methodology gloss over the implementation. This is particularly important with modern Bayesian methods applied to marketing problems. The complexity of the models and the dimensionality of the data can render some methods impractical. MCMC methods can be theoretically valid but of little practical value. Computations lasting more than a day can be required for adequate inference due to high autocorrelation and slow computation of an iteration of the chain.

If a method takes more than 3 or 4 hours of computing time on standard equipment, we deem it impractical in the sense that most investigators are unwilling to wait much longer than this for results. However, what is practical depends not only on the speed of computing equipment but on the quality of the implementation. The good news is that in 2024, even the most pedestrian computing equipment is capable of truly impressive computations, unthinkable at the beginning of the MCMC revolution in the late 1980s and early 1990s. Achieving the theoretical capabilities of the latest CPU chip may require much specialized programming, use of optimized BLAS libraries and the use of a low-level language such as C or FORTRAN. Most investigators are not willing to make this investment unless their primary focus is on the development of methodology. Thus, we view a method as "practical" if it can be computed in a relatively high-level computing environment which mimics as closely as possible the mathematical formulas for the methods and models. For even wider dissemination of our methods, some sort of pre-packaged set of methods and models is also required.

For these reasons, we decided to program the models and methods of this book in the R language. In addition, we provide a web site for the book which provides further data and code for models discussed in the case studies. R is free, widely accepted in the statistical community, and offers much of the basic functionality needed and support for optimized matrix operations. Originally, our supporting code was written primarily in R with only a few functions translated into C. Since version 3.0 of `bayesm`, we have converted all computations into C++ and use the Armadillo matrix class. R is only used as a wrapper for these functions and to provide rudimentary checks on the validity of arguments. We have not implemented the use of parallelization or GPUs to enhance the speed of execution for our code. It is entirely possible that very large improvements in speed could be achieved with such improvements. Unfortunately, there are some issues at present that make it difficult to implement these approaches in a way that is transparent to hardware and software architecture and will supports the basic UNIX, Windows, and MacOS set of machines.

R

CPU speed is not the only resource that is important in computing. Memory is another resource which can be a bottleneck. Our view is that memory is so cheap that we do not want to modify our code to deal with memory constraints. All of our programs are designed to work entirely in memory. All of our applications use less than 10 GBs of memory.

Our experiences coding and profiling the applications in this book have changed our views on statistical computing. We were raised to respect minor changes in the speed of computations via various tricks and optimization of basic linear algebra operations. When we started to profile our code, we realized that, to a first approximation, linear algebra is free. The mainstay of Bayesian computations is the Cholesky root. These are virtually free on modern equipment (for example, one can compute the Cholesky root of 1000×1000 matrices at the rate of at least 200 per minute on standard-issue laptop computers). We found conversions from vectors to matrices and other "minor" operations to be more computationally demanding. Minimizing the number of matrix decompositions or taking advantage of the special structure of the matrices involved often has only minor impact. Optimization frequently involves little more than avoiding loops over the observations in the data set.

Computing also has an important impact on those who wish to learn from this book. We recognized, from the start, that our audience may be quite diverse. It is easy to impose

a relatively minimal requirement regarding the level of knowledge of Bayesian statistics. It is harder to craft a set of programs which can be useful to readers with differing computing expertise and time to invest in computing. We decided that a two-pronged attack was necessary: 1. for those who want to use models pretty much "off-the-shelf," we have developed an R package to implement most of the models developed in the book; and 2. for those who want to learn via programming and who wish to extend the methods and models, we provide detailed code and examples for each of the chapters
R of the book and for each of the case studies. Our R package, `bayesm`, is available on the Comprehensive R Archive Network (CRAN, google "R language" for the URL). `bayesm` implements all of the models and methods discussed in the first seven chapters (see Appendix A for an introduction to R and `bayesm`). The book's website provides documented code, data sets, and additional information for those who which to adapt our models and methods.

We provide this code and examples with some trepidation. In some sense, those who really want to learn this material intimately will want to write their own code from scratch, using only some of our basic functions. We hope that providing the "answers" to the problem will not discourage study. Rather, we hope many of our readers will take our code as a base to improve on. We expect to see much innovation and improvement on what we think is a solid base.

ACKNOWLEDGMENTS

We owe a tremendous debt to our teachers, Dennis Lindley and Arnold Zellner, who impressed upon us the value and power of the Bayesian approach. Their enthusiasm is infectious. Our students (including Andrew Ainslie, Neeraj Arora, Peter Boatwright, Yancy Edwards, Tim Gilbride, Lichung Jen, Ling-Jing Kao, Jaehwan Kim, Alan Montgomery, Sandeep Rao, and Sha Yang) have been great collaborators as well as patient listeners as we have struggled to articulate this program of research. Junhong Chu read the manuscript very carefully and rooted out numerous errors and expositional problems. George Hochwarter provided invaluable assistance in converting the first edition manuscript into LaTex. Allenby would like to thank Vijay Bhargava for his encouragement in both personal and professional life. Rossi thanks the Anderson School of Management at UCLA for support. Misra thanks the Booth School of Business and the Kilts Center for Marketing at the University of Chicago for support.

2
Bayesian Essentials

Abstract

This chapter provides a self-contained introduction to Bayesian Inference. For those who need a refresher in distribution theory, Section 2.1 provides an introduction to marginal, joint, and conditional distributions and their associated densities. We then develop the basics of Bayesian inference, discuss the role of subjective probability and priors and provide some of the most compelling arguments for adopting the Bayesian point of view. Regression models (both univariate and multivariate) are considered along with their associated natural conjugate priors. Asymptotic approximations and Importance Sampling are introduced as methods for non-conjugate models. Finally, a simulation primer for the basic distributions/models in Bayesian Inference is provided. Those who want a basic introduction to Bayesian inference without many details should concentrate on Sections 2.2–2.6 and Section 2.10.1.

2.1 ESSENTIAL CONCEPTS FROM DISTRIBUTION THEORY

Bayesian inference relies heavily on probability theory and, in particular, distributional theory. This section provides a review of basic distributional theory with examples designed to be relevant to Bayesian applications.

A basic starting point for probability theory is a discrete random variable, X. X can take on a discrete number of values, each with some probability. The classic example would be a Bernoulli random variable. $X = 1$ with probability p and 0 with probability $p - 1$. X denotes some event such as whether a company will sell a product tomorrow. p represents the probability of a sale. For now, let us set aside the question of whether this probability can represent a long run frequency or whether it represents a subjective probability (note: it is hard to understand the long-run frequency argument for this example since it requires us to imagine an infinite number of "other-worlds" for the event of a sale tomorrow). We can easily extend this example to the number of units sold tomorrow. Then X is still discrete but can take on the values $0, 1, 2, \ldots, m$ with probabilities, $p_0, p_1, \ldots, p_m$. X now has a nontrivial probability distribution. With knowledge of this distribution, we can answer any question such as the probability that there will be at

Bayesian Statistics and Marketing, Second Edition. Peter E. Rossi, Greg M. Allenby, and Sanjog Misra

least one sale tomorrow, the probability that there will be between 1 and 10 sales, etc. In general, we can compute the probability that sales will be in any set simply by summing over the probabilities of the elements in the set.

$$\Pr(X \in A) = \sum_{x \in A} p_x \tag{2.1.1}$$

We can also compute the *expectation* of the number of units sold tomorrow as the average over the probability distribution.

$$E[X] = \sum_{i=0}^{m} i p_i \tag{2.1.2}$$

If we are looking at aggregate sales of a popular consumer product, we might approximate sales as a *continuous* random variable that can take on any nonnegative real number. For this situation, we must summarize the probability distribution of X by a probability density. A density function is a *rate* function that tells us the probability per volume or unit of X. X has a density function, $p_X(x)$; p_X is a positive-valued function that integrates to one. The probability that X takes on any set of values we must integrate $p_X(\,)$ over this set.

$$\Pr(X \in A) = \int_A p_X(x|\theta)\, dx \tag{2.1.3}$$

This is very much the analogue of the discrete sum in (2.1.1). The sense in which p is a rate function is that the probability that $X \in (x_0, x_0 + dx)$ is approximately $p_X(x_0)\, dx$. Thus, the probability density function, $p_X(\,)$, plays the same role as the discrete probabilities (sometimes called probability mass function) in the discrete case. We can easily find the expectation of any function of X by computing the appropriate integral.

$$E[f(X)] = \int f(x)\, p(x|\theta)\, dx \tag{2.1.4}$$

In many situations, we will want to consider the *joint* distribution of two or more random variables, both of which are continuous. For example, we might consider the joint distribution of sales tomorrow in two different markets. Let X denote the sales in market A and Y denote the sales in market B. For this situation, there is a bivariate density function, $p_{X,Y}(x, y)$. This density gives the probability rate per unit of area in the plane. That is, the probability that both $X \in (x_0, x_0 + dx)$ and $Y \in (y_0, y_0 + dy)$ is approximately, $p_{X,Y}(x_0, y_0)\, dxdy$. With the joint density, we compute the probability of any set of (X, Y) values. For example, we can compute the probability that both X and Y are positive. This is the area of under the density for the positive orthant.

$$\Pr(X > 0 \text{ and } Y > 0) = \int_0^\infty \int_0^\infty p_{X,Y}(x, y)\, dxdy \tag{2.1.5}$$

For example, the multinomial probit model, considered in Chapter 4, has choice probabilities defined the integrals of a multivariate normal density over various cones. If $p_{X,Y}(\,)$ is a bivariate normal density, then (2.1.5) is one such equation.

Given the joint density, we can also compute the *marginal* densities of each of the variables X and Y. That is to say, if we know everything thing about the joint distribution, we certainly know everything about the marginal distribution. The way to think of this is

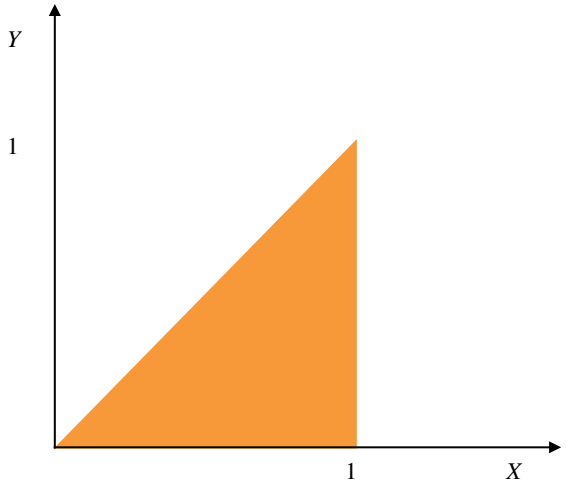


Figure 2.1 Support for the example of a bivariate distribution

via simulation. Suppose we were able to simulate from the joint distribution. If we look at the simulated distribution of either X or Y alone, we have simulated the marginal distribution.

To find the marginal density of X, we must average the *joint* density over all possible values of Y.

$$p_X(x) = \int p_{X,Y}\left(x, y\right) dy \tag{2.1.6}$$

A simple example will help make this idea clear. Suppose X, Y are uniformly distributed over the triangle, $\{X, Y : 0 < X < 1 \text{ and } Y < X\}$, depicted in Figure 2.1. A uniform distribution means that the density is constant over the shaded triangle. The area of this triangle is $\frac{1}{2}$ so this means that the density must be 2 in order to insure that the joint density integrates to 1.

$$\int_0^1 \int_y^1 p_{X,Y}\left(x, y\right) dxdy = \int_0^1 \int_y^1 2\, dxdy = \int_0^1 \left(2x|_y^1\right) dy$$
$$= \int_0^1 \left(2 - 2y\right) dy = 2y - y^2|_0^1 = 1$$

This means that the joint density is a surface over the triangle with height 2.

We can use (2.1.6) to find the marginal distribution of X by integrating out Y.

$$p_X(x) = \int p_{X,Y}\left(x, y\right) dy = \int_0^x 2\, dy = 2y|_0^x = 2x$$

Thus, the marginal distribution of X is not uniform! The density increases as x increases toward 1. The marginal density of Y can easily be found to be of the "reverse" shape, $p_Y\left(y\right) = 2 - 2y$. This makes intuitive sense as the joint density is defined over the "widest" area with X near one and with Y near 0.

We can also define the concept of a conditional distribution and conditional density. If X, Y have a joint distribution, we can ask what is the conditional distribution of Y given

$X = x$? If X, Y are continuous random variables, then the conditional distribution of Y given $X = x$ is also a continuous random variable. The conditional density of $Y \mid X$ can be derived from the marginal and joint densities (the Borel paradox not withstanding).

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \tag{2.1.7}$$

The argument of the conditional density on the left hand side of (2.1.7) is written $y|x$ to emphasize that there is a different density for every value of the conditioning argument x. We note that the conditional density is proportional to the joint! The marginal only serves to get the right normalization.

Let's return to our simple example. The conditional distribution of $Y|X = x$ is simply a slice of the joint density along a vertical line at the point x. This is clearly uniform but only extends from 0 to x. We can use (2.1.7) to get the right normalization.

$$p_{Y|X}(y|x) = \frac{2}{2x}; \;\; y \in (0,x)$$

Thus, if $x = 1$, then the density is uniform over (0,1) with height 1. The dependence between X and Y is only evidenced by the fact the range of Y is restricted by the value of x.

In many statistics courses, we are taught that correlation is a measure of the dependence between two random variables. This stems from the bivariate normal distribution that uses correlation to drive the shape of the joint density. Let's start with two independent standard normal random variables, Z and W. This means that their joint density factors (this is because of the product rule for independent events).

$$p_{Z,W}(z,w) = p_Z(z)\, p_W(w) \tag{2.1.8}$$

Each of the standard normal densities is given by:

$$p_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{2.1.9}$$

If we create X and Y by an appropriate linear combination of Z and W, we can create correlated or dependent random variables.

$$\begin{aligned} X &= Z \\ Y &= \rho Z + \sqrt{\left(1-\rho^2\right)} W \end{aligned}$$

X and Y have a correlated bivariate normal density with correlation coefficient ρ.

$$p_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[x^2 - 2\rho xy + y^2\right]\right\} \tag{2.1.10}$$

It is possible to show that $cov(X,Y) = E[XY] = \iint xy p_{X,Y}(x,y)\, dxdy$ is ρ. Both X, Y have marginal distributions that are standard normal and conditional distributions which are also normal but with a mean that depends on the conditioning argument.

$$X \sim N(0,1);\; Y \sim N(0,1);\; Y|X = x \sim N\left(\rho x, \left(1-\rho^2\right)\right)$$

We will return to this example when we consider methods of simulation from the bivariate and multivariate normal distributions. We will also consider this situation when introducing the Gibbs Sampler in Chapter 3.

2.2 THE GOAL OF INFERENCE AND BAYES THEOREM

The goal of statistical inference is to use information to make inferences about unknown quantities. One important source of information is data but there is an undeniable role for non data-based information. Information can also come from theories of behavior (such as the information that, properly defined, demand curves slope downward). Information can also come from "subjective" views that there is a structure underlying the unknowns. For example, in situations with large numbers of different sets of parameters, an assumption that the parameters sets "cluster" or that they are drawn from some common distribution is often used in modeling. Less controversial might be the statement that we expect key quantities to be finite or even in some range (such as a price elasticity is not expected to be less than -50). Information can also be derived from prior analyses of other data, including data that is only loosely related to the dataset under investigation.

An unknown quantity is a generic term referring to any value not known to the investigator. Certainly, parameters can be considered unknown since these are purely abstractions that index a class of models. In situations in which decisions are made, the unknown quantities can include the, as yet unrealized, outcomes of marketing actions. Even in a passive environment, predictions of "future" outcomes are properly regarded as unknowns. There should be no distinction between a parameter and an unknown such as an unrealized outcome in the sense that the system of inference should treat each symmetrically.

Our goal, then, is to make inferences regarding unknown quantities *given* the information available. We have concluded that the information available can be partitioned into information obtained from the data as well as other information obtained independently or *prior* to the data. Bayesian inference utilizes probability statements as the basis for inference. What this means is that our goal is to make probability statements about unknown quantities *conditional* on the sample and prior information.

In order to utilize the elegant apparatus of conditional probability, we must encode the prior information as a probability distribution. This requires the view that probability can represent subjective beliefs and is not some sort of long run frequency. There is much discussion in the statistics and probability theory literature as to whether or not this is a reasonable thing to do. We take a somewhat more practical view – there are many kinds of non-data-based information to be incorporated into our analysis. A subjective interpretation of probability is a practical necessity rather than a philosophical curiosity.

It should be noted that there are several paths which lead to the conclusion that Bayesian inference is a sensible system of inference. Some start with the view that decision makers are expected utility maximizers. In this world, decision makers must be "coherent" or act in accordance with Bayes theorem in order to avoid exposing themselves to sure losses. Others start with the view that the fundamental primitive is not utility but subjective probability. Still others adhere to the view that the likelihood principle (Section 2.3) more or less forces you to adopt the Bayesian form of inference. We are more of the subjectivist stripe but we hope to convince the reader, by example, that there is tremendous practical value to the Bayesian approach.

2.2.1 Bayes Theorem

Denote the set of unknowns as θ. Our prior beliefs are expressed as a probability distribution, $p(\theta)$. $p(\bullet)$ is a generic notation for the appropriate density. In most cases, this

represents a density with respect to standard Lebesgue measure but it can also represent a probability mass function for discrete parameter spaces or a mixed continuous-discrete measure. The information provided by the data is introduced via the probability distribution for the data, $p(D|\theta)$, where "D" denotes the observable data. In some classical approaches, modeling is the art of choosing appropriate probability models for the data. In the Bayesian paradigm, the model for prior information is also important. Much of the work in Bayesian statistics is focused on developing a rich class of models to express prior information and devices to induce priors on high dimensional spaces. In our view, the prior is very important and often receives insufficient attention.

To deliver on the goal of inference, we must combine the prior and likelihood to produce the distribution of the observables conditional on the data and the prior. Bayes Theorem is nothing more than an application of standard conditional probability to this problem.

$$p(\theta|\, D) = \frac{p(D, \theta)}{p(D)} = \frac{p(D|\, \theta)\, p(\theta)}{p(D)} \tag{2.2.1}$$

$p(\theta|\, D)$ is called the *posterior* distribution and reflects the combined data and prior information. (2.2.1) is often expressed using the likelihood function. Given D, any function that is proportional to $p(D|\theta)$ is call the "likelihood," $\ell(\theta)$. The shape of the posterior is determined entirely by the likelihood and prior in the numerator of (2.2.1) and this is often emphasized by rewriting the equation.

$$p(\theta|\, D) \propto \ell(\theta)\, p(\theta) \tag{2.2.2}$$

If $\ell(\theta) = p(D|\theta)$, then the constant of proportionality is the marginal distribution of the data, $p(D) = \int p(D, \theta)\ d\theta = \int p(D|\theta)\ p(\theta)\, d\theta$. Of course, we are assuming here that this normalizing constant exists. If $p(\theta)$ represents a proper distribution (i.e., it integrates to one), then this integral will likely exist. With improper priors, it will be necessary to show that the integral exists, which will involve the tail behavior of the likelihood functions.

2.3 CONDITIONING AND THE LIKELIHOOD PRINCIPLE

The likelihood principle states that the likelihood function, $\ell(\theta)$, contains all relevant information from the data. Two samples (not necessarily even from the same "experiment" or method of sampling/observation) have equivalent information regarding θ if their likelihoods are proportional (see Berger and Wolpert [1984] for extensive discussion and derivation of the LP from conditioning and sufficiency principles). The likelihood principle, by itself, is not sufficient to build a method of inference but should be regarded as a minimum requirement of any viable form of inference. This is a controversial point of view for anyone familiar with the modern econometric literature. Much of this literature is devoted to methods that do not obey the likelihood principle. For example, the phenomenal success of estimators based on the Generalized Method of Moments procedure is driven by the ease of implementing these estimators even though, in most instances, GMM estimators violate the likelihood principle.

Adherence to the likelihood principle means that inferences are *conditional* on the observed data as the likelihood function is parameterized by the data. This is worth

contrasting to any sampling-based approach to inference. In the sampling literature, inference is conducted by examining the sampling distribution of some estimator of θ, $\hat{\theta} = f(D)$. Some sort of sampling experiment[1] results in a distribution of D and, therefore, the estimator is viewed as a random variable. The sampling distribution of the estimator summarizes the properties of the estimator *prior* to observing the data. As such, it is irrelevant to making inferences given the data we actually observe. For any finite sample, this distinction is extremely important. One must conclude that, given our goal for inference, sampling distributions are simply not useful.

While sampling theory does not seem to deliver on the inference problem, it is possible to argue that it is relevant to the choice of estimating procedures. Bayesian inference procedures are simply one among many possible methods of deriving estimators for a given problem. Sampling properties are relevant to choice of procedures before the data is observed. As we will see in Section 2.6, there is an important sense in which one need never look farther than Bayes estimators even if the sole criterion is the sampling properties of the estimator.

2.4 PREDICTION AND BAYES

One of the appeals of the Bayesian approach is that all unknowns are treated the same. Prediction is defined as making probability statements about the distribution of as yet unobserved data, denoted by D_f. The only real distinction between "parameters" and unobserved data is that D_f is potentially observable.

$$p\left(D_f|\, D\right) = \int p\left(D_f, \theta|D\right)\, d\theta = \int p\left(D_f|\theta, D\right)\, p(\theta|D)\, d\theta \tag{2.4.1}$$

(2.4.1) defines the "predictive" distribution of D_f given the observed data. In many cases, we assume that D and D_f are independent, conditional on θ. In this case, the predictive distribution simplifies.

$$p\left(D_f|D\right) = \int p\left(D_f|\,\theta\right)\, p(\theta\,|D)\, d\theta \tag{2.4.2}$$

In (2.4.2), we average the likelihood for the unobserved data over the posterior of θ. This averaging properly accounts for uncertainty in θ when forming predictive statements about D_f.

2.5 SUMMARIZING THE POSTERIOR

For any problem of practical interest, the posterior distribution is a high dimensional object. Therefore, summaries of the posterior play an important role in Bayesian

[1] In the standard treatment, the sampling experiment consists of draws from the probability model for the data used in the likelihood. However, many other experiments are possible including samples experiments which involve additional assumptions regarding the data generation process.

statistics. Most schooled in classical statistical approaches are accustomed to reporting parameter estimates and standard errors. The Bayesian analogue of this practice is to report moments of the marginal distributions of parameters such as the posterior mean and posterior standard deviations. It is far more useful and informative to produce the marginal distributions of parameters or relevant functions of parameters as the output of the analysis. Simulation methods are ideally suited for this. If we can simulate from the posterior distribution of the parameters and other unknowns, then we can simply construct the marginal of any function of interest. Typically, we describe these marginals graphically. As these distributions are often very nonnormal, the mean and standard deviations are not particularly useful. One major purpose of this book is to introduce a set of useful simulation tools to achieve this goal of simulating from the posterior distribution.

Prior to the advent of powerful simulation methods, attention focused on the evaluation of specific integrals of the posterior distribution as a way of summarizing this high dimensional object. The general problem can be written as finding the posterior expectation of a function of θ. (We note that marginal posteriors, moments, quantiles, and probability of intervals are further examples of expectations of functions as in (2.5.1) with suitably defined h). For any interesting problem, only the un-normalized posterior, $\ell(\theta)\, p(\theta)$ is available so that two integrals must be performed to obtain the posterior expectation of $h(\theta)$

$$E_{\theta|D}\left[h(\theta)\right] = \int h(\theta)\ p(\theta|D)\ d\theta = \frac{\int h(\theta)\, \ell(\theta)\, p(\theta)\, d\theta}{\int \ell(\theta)\, p(\theta)\, d\theta} \tag{2.5.1}$$

For many years, only problems for which the integrals in (2.5.1) could be performed analytically were analyzed by Bayesians. Obviously, this restricts the set of priors and likelihoods to a very small set that produces posteriors of known distributional form and for which these integrals can be evaluated analytically. One approach would be to take various asymptotic approximations to these integrals. We will discuss the Laplace approximation method in Section 2.10. Unless these asymptotic approximations can be shown to be accurate, we should be very cautious about using them. In contrast, much of the econometrics and statistics literature uses asymptotic approximations to the sampling distributions of estimators and test statistics without investigating accuracy. In marketing problems, the combination of small amounts of sample information per parameter and the discrete nature of the data makes it very risky to use asymptotic approximations. Fortunately, we do not have to rely on asymptotic approximations in modern Bayesian inference.

2.6 DECISION THEORY, RISK, AND THE SAMPLING PROPERTIES OF BAYES ESTIMATORS

We started our discussion by posing the problem of obtaining a system of inference appropriate for marketing problems. We could just as well have started on the most general level – finding an appropriate framework for making decisions of any kind. Parameter estimation is only one of many such decisions that occur under uncertainty.

The general problem considered in decision theory is to search among possible actions for the action that minimizes expected loss. The loss function, $L(a, \theta)$, associates a loss

with a state of nature (θ) and an action a. In Chapter 6, loss functions are derived for marketing actions from the profit function of the firm. We choose a decision that performs well, on average, where the averaging is taken across the posterior distribution of states of nature.

$$\min_{a} \left\{ \bar{L}(a) = E_{\theta|D}[L(a,\theta)] = \int L(a,\theta)\, p(\theta|D)\, d\theta \right\} \tag{2.6.1}$$

In Chapter 6, we will explore the implications of decision theory for optimal marketing decisions and valuing of information sets. At this point it is important to note that (2.6.1) involves the entire posterior distribution and not just the posterior mean. With nonlinear loss functions, uncertainty or spread is just as important as location.

A special case of (2.6.1) is the estimation problem. If the action is the estimator and the state of nature is the unknowns to be estimated, then Bayesian decision theory produces a Bayes estimator. Typically, a symmetric function such as squared error or absolute error is used for loss. This defines the estimation problem as

$$\min_{\hat{\theta}} \left\{ L\left(\hat{\theta}\right) = E_{\theta|D}\left[L\left(\hat{\theta},\theta\right)\right] \right\}. \tag{2.6.2}$$

For squared error loss, the optimal choice of estimator is the posterior mean.

$$\hat{\theta}_{Bayes} = E[\theta|D] = f(D|\tau) \tag{2.6.3}$$

Here τ is the prior hyper-parameter vector (if any). If the prior is a parametric family of distributions, then the prior hyper-parameters are the parameters that describe this family. For example, if the prior is a normal distribution, then the prior mean and prior variance comprise the prior hyper-parameters.

What are the sampling properties of the Bayes estimator and how do these compare to those of other competing general purpose estimation procedures such as Maximum Likelihood? Recall the sampling properties are derived from the fact that the estimator is a function of the data and therefore is a random variable whose distribution is inherited from the sampling distribution of the data. We can use the same loss function to define the "risk" associated with an estimator, $\hat{\theta}$, as

$$r_{\hat{\theta}}(\theta) = E_{D|\theta}\left[L\left(\hat{\theta},\theta\right)\right] = \int L\left(\hat{\theta}(D),\theta\right) p(D|\theta)\, dD \tag{2.6.4}$$

Note that the risk function for an estimator is a function of θ. That is, we have a different "risk" at every point in the parameter space.

An estimator is said to be *admissible* if there exists no other estimator with a risk function that is less than or equal to the risk of the estimator in question. That is, we cannot find another estimator that does better (or at least as well, as measured by risk, for every point in the parameter space.[2] Define expected risk, $E[r(\theta)] = E_{\theta}\left[E_{D|\theta}\left[L\left(\hat{\theta},\theta\right)\right]\right]$. The outer expectation on the right hand side is taken with respect to the prior distribution of θ. With a proper prior that has support over the entire parameter space, we can apply Fubini's theorem and interchange the order of integration and show that Bayes estimators

[2] Obviously if we have a continuous parameter space, we have to be a little more careful but we leave those niceties for those more mathematically inclined.

have the property of minimizing expected risk and, therefore, are admissible.

$$\begin{aligned} E[r(\theta)] = \quad & E_\theta\left[E_{D|\theta}\left[L\left(\hat{\theta},\theta\right)\right]\right] = \iint L\left(\hat{\theta}(D),\theta\right)\, p(D|\theta)\, p(\theta)\, dD d\theta \\ = \quad & E_D\left[E_{\theta|D}\left[L\left(\hat{\theta},\theta\right)\right]\right] = \iint L\left(\hat{\theta}(D),\theta\right)\, p(\theta|D)\, p(D)\, dD d\theta \end{aligned} \tag{2.6.5}$$

The complete class theorem (see Berger [1985], Chapter 8) says even more – all admissible estimators are Bayes estimators. This provides a certain level of comfort and moral superiority but little practical guidance. There can be estimators that outperform Bayes estimators in certain regions of, but not all, of the parameter space. Bayes estimators perform very well if you are in the region of the parameter space you expect to be in as defined by your prior. These results on admissibility also don't provide any guidance as to how to choose among infinite number of Bayes estimators that are equivalent from the point of view of admissibility.

Another useful question to ask is what is the relationship between standard classical estimators such as the MLE and Bayes estimators? At least the MLE obeys the likelihood principle. In general, the MLE is not admissible so there can be no exact sample relationship. However, Bayes estimators are consistent, asymptotically normal and efficient as long as mild regularity conditions[3] hold and the prior is nondogmatic in the sense of giving support to the entire parameter space. The asymptotic "duality" between Bayes estimators and the MLE stems from the asymptotic behavior of the posterior distribution. As n increases, the posterior concentrates more and more mass in the vicinity of the "true" value of θ. The likelihood term dominates the prior and the prior becomes more and more uniform in appearance in the region in which the likelihood is concentrating. Thus, the prior has no asymptotic influence and the posterior starts to look more and more normal.

$$p(\theta|\, D) \dot{\sim} N\left(\hat{\theta}_{MLE}, \left[-H_{\theta=\hat{\theta}_{MLE}}\right]^{-1}\right) \tag{2.6.6}$$

H_θ is the Hessian of the log-likelihood function. The very fact that, for asymptotics, the prior doesn't matter (other than its support) should be reason enough to abandon this method of analysis in favor of more powerful techniques.

2.7 IDENTIFICATION AND BAYESIAN INFERENCE

The set of models is only limited by the imagination of the investigator and the computational demands of the model and inference method. In marketing problems, we can easily write down a model that is very complex and may make extraordinary demands of data. A problem of identification is defined as the situation in which there is a set of different parameter values that give rise to the same distribution for the data. This set of parameter

[3] It should be noted that as the MLE is based on a maximum of a function while the Bayes estimator is based on an average, the conditions for asymptotic normality are different for the MLE than for the Bayes estimator. But both from a practical (i.e., computational) and theoretical perspective, averages behave more regularly than maxima.