Bayesian Statistics and Marketing

Peter E. Rossi
University of Chicago,USA

Greg M. Allenby
Ohio State University, USA

Robert McCulloch
University of Chicago, USA

John Wiley & Sons, Ltd

Bayesian Statistics and Marketing

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Peter Bloomfield, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels;*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

A complete list of the titles in this series appears at the end of this volume.

Bayesian Statistics and Marketing

Peter E. Rossi
University of Chicago,USA

Greg M. Allenby
Ohio State University, USA

Robert McCulloch
University of Chicago, USA

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Reprinted with corrections August 2006

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data:

Rossi, Peter E. (Peter Eric), 1955-
Bayesian, statistics and marketing / Peter Rossi and Greg Allenby, Rob McCulloch.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-470863-67-1 (HB)
ISBN-10: 0-470863-67-6 (HB)
1. Marketing research – Mathematical models. 2. Marketing – Mathematical models. 3. Bayesian statistical decision theory. I. Allenby, Greg M. (Greg Martin), 1956- II. McCulloch, Robert E. (Robert Edward) III. Title.
HF5415.2.R675 2005
658.8′3′015118 – dc22
2005016418

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470863-67-1 (HB)
ISBN-10: 0-470863-67-6 (HB)

Typeset in 10/12pt Galliard by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

To our wives –
Laurie, Tricia, and Nancy

for our children –
Ben and Emily, Kate and Mark, Kate and Stephen

and with thanks to our parents –
Alice and Peter, Marilyn and Stan, Ona and Ernest

Contents

1
Introduction

The past ten 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 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.

Until the mid-1980s, Bayesian methods appeared to be 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 match 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

Bayesian Statistics and Marketing P. E. Rossi, G. M. Allenby and R. McCulloch

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 regard the customer-level parameters not as nuisance parameters but 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 who find highest value in the offering's attributes and those who 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 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 instance 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 the problem is the conditional likelihood for the unit-level behavior. We condition on unit-specific parameters which 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 the 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 which 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,

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

If we observe more than one observation per consumer, it is natural to consider a model which 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 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:

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

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 (1.2.2). 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]

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

Here $\varphi(\cdot)$ 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

[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.

common parameters that drive the distribution of βs across consumers but also the individual βs.

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:

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

Ω represents the information available about the distribution of the outcomes resulting from marketing actions. Clearly, information about 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 be practical ways of implementing this system of inference. By 'practical' we mean computable for problems of the size which practitioners in marketing encounter.

Throughout 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.

1.3 BENEFITS AND COSTS OF THE BAYESIAN APPROACH

At the beginning of Chapter 2, we outline the basics of the Bayesian approach to inference and decision-making. There are really no other approaches which can provide a unified treatment of inference and decision as well as properly accounting for parameter and model uncertainty. However compelling the logic 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 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: (i) formulation of a prior; (ii) requirement of a likelihood function; and (iii) 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 on 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-based information, priors in a fairly broad range will result in much the same a posteriori inferences. However, in any problem in which the data-based information to 'parameters' ratio is low, priors will matter. In models with unit-level parameters, there is often relatively little data-based information, so that it is vital that the system of inference incorporate 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 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 misspecification 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 a way of guarding against misspecification. 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 which 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 also 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 at 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 at 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 which 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 much attention to models for discrete data as much disaggregate marketing data involves 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 exogeneous variables. In marketing applications, the marketing mix variables and sales are jointly determined given a set of exogeneous demand or cost shifters.

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 which involve extensions or elaborations of the material covered in the first seven chapters. Each of the case studies has 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 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 that we have

not 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 also on the quality of the implementation. The good news is that in 2005 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 prepackaged 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 website 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. We have taken the philosophy advocated by Chambers and others that one should code in R, profile the code, and, if necessary, rewrite in a lower-level language such as C. This philosophy has produced over 4000 lines of R code to implement the models and methods in this book and less than 500 lines of C and C++ code. We have been impressed by the speed of R running on standard computing equipment. This is a testimony to the speed of modern CPUs as well as the hard work of the R development team and contributors. In many instances, we can achieve more than adequate speed without any low-level code. We do not claim to have come anywhere near the theoretical speed possible for these applications. The gap between what is theoretically possible and what is achievable in R is only important if you are faced with a computing bottleneck.

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 1 GB of memory. Given that 512 MB is now relatively standard for Windows machines, we think this is reasonable as much as it might raise eyebrows among those who were brought up in a memory-poor world.

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-prong attack was necessary: First, 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 second, 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 of the book and for each of the case studies.

R 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 Chapters 1–7 (see Appendix B for more information on `bayesm` and Appendix A for an introduction to R). The book's website, www.wiley.com/go/bsm, provides documented code, data sets and additional information for those who wish to adapt our models and methods. Throughout this book, a boldface "R" in the left margin indicates a reference to `bayesm`.

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.

ACKNOWLEDGEMENTS

We owe a tremendous debt to our teachers, Seymour Geisser, 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 and Maria Ana Vitorino read the manuscript very carefully and rooted out numerous errors and expositional problems. Greg would like to thank Vijay Bhargava for his encouragement in both personal and professional life.

McCulloch and Rossi gratefully acknowledge support from the Graduate School of Business, University of Chicago. Rossi also thanks the James M. Kilts Center for Marketing, Graduate School of Business, University of Chicago for support. Allenby gratefully acknowledges support from the Fisher College of Business, Ohio State University.

2
Bayesian Essentials

Using this Chapter

This chapter provides a self-contained introduction to Bayesian inference. For those who need a refresher in distribution theory, Section 2.0 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 nonconjugate models. Finally, a simulation primer for the basic distributions/models in Bayesian inference is provided. Those who want an introduction to Bayesian inference without many details should concentrate on Sections 2.1–2.5 and Section 2.8.1.

2.0 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 countable number of values, each with some probability. The classic example would be a Bernoulli random variable, where X takes the value 1 with probability p and 0 with probability $1 - p$. 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 least one sale tomorrow, the probability that there

Bayesian Statistics and Marketing P. E. Rossi, G. M. Allenby and R. McCulloch

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:

$$\Pr(X \in A) = \sum_{x \in A} p_x. \tag{2.0.1}$$

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

$$E[X] = \sum_{i=0}^{m} i p_i. \tag{2.0.2}$$

If we are looking at aggregate sales of a popular consumer product, we might approximate sales as a *continuous* random variable which 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 which 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 which integrates to one. To find the probability that X takes on any set of values we must integrate $p_X(\bullet)$ over this set:

$$\Pr(X \in A) = \int_A p_X(x|\theta)\, \mathrm{d}x. \tag{2.0.3}$$

This is very much the analog of the discrete sum in (2.0.1). The sense in which p is a rate function is that the probability that $X \in (x_0, x_0 + \mathrm{d}x)$ is approximately $p_X(x_0)\, \mathrm{d}x$. Thus, the probability density function, $p_X(\bullet)$, plays the same role as the discrete probability (sometimes called the probability mass function) in the discrete case. We can easily find the expectation of any function of X by computing the appropriate integral:

$$E[f(X)] = \int f(x) p(x|\theta)\, \mathrm{d}x. \tag{2.0.4}$$

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 $X \in (x_0, x_0 + \mathrm{d}x)$ and $Y \in (y_0, y_0 + \mathrm{d}y)$ is approximately $p_{X,Y}(x_0, y_0)\, \mathrm{d}x\, \mathrm{d}y$. 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:

$$\Pr(X > 0 \text{ and } Y > 0) = \int_0^\infty \int_0^\infty p_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y. \tag{2.0.5}$$

For example, the multinomial probit model, considered in Chapter 4, has choice probabilities defined by the integrals of a multivariate normal density over various cones. If $p_{X,Y}(\bullet, \bullet)$ is a bivariate normal density, then (2.0.5) 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 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:

$$p_X(x) = \int p_{X,Y}(x, y)\, \mathrm{d}y. \tag{2.0.6}$$

A simple example will help make this idea clear. Suppose X, Y are uniformly distributed over the triangle $\{X, Y : 0 < Y < 1 \text{ and } Y < X < 1\}$, 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 ensure that the joint density integrates to 1:

$$\begin{aligned}\int_0^1 \int_y^1 p_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y &= \int_0^1 \int_y^1 2\, \mathrm{d}x\, \mathrm{d}y = \int_0^1 (2x\big|_y^1)\, \mathrm{d}y \\ &= \int_0^1 (2 - 2y)\, \mathrm{d}y = (2y - y^2)\big|_0^1 = 1\end{aligned}$$

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

We can use (2.0.6) to find the marginal distribution of X by integrating out Y:

$$p_X(x) = \int p_{X,Y}(x, y)\, \mathrm{d}y = \int_0^x 2\, \mathrm{d}y = 2y\big|_0^x = 2x.$$

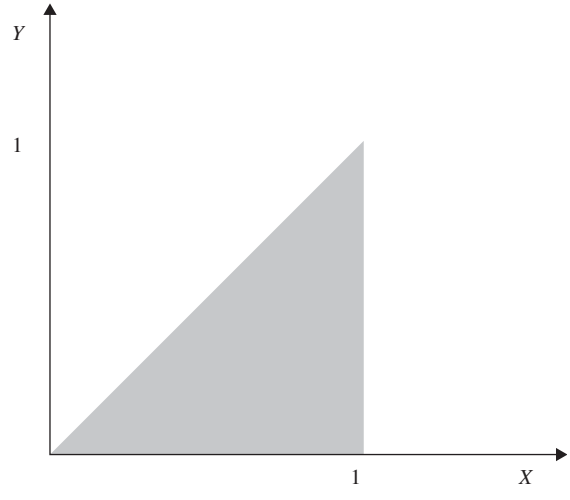

Figure 2.1 Support for the example of a bivariate distribution

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, $p_Y(y) = 2 - 2y$. This makes intuitive sense as the joint density is defined over the 'widest' area with X near 1 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 $Y|X$ can be derived from the marginal and joint densities:[1]

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}. \tag{2.0.7}$$

The argument of the conditional density on the left-hand side of (2.0.7) 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 us 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 (2.0.7) to get the right normalization.

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

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 that 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 which uses correlation to drive the shape of the joint density. Let us start with two independent standard normal random variables, Z and W. This means that their joint density factorizes:

$$p_{Z,W}(z, w) = p_Z(z)p_W(w) \tag{2.0.8}$$

(this is because of the product rule for independent events). Each of the standard normal densities is given by

$$p_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{2.0.9}$$

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

$$X = Z, \qquad Y = \rho Z + \sqrt{(1 - \rho^2)}\, W.$$

[1] The Borel paradox not withstanding.

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

$$p_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[x^2 - 2\rho xy + y^2\right]\right\}. \tag{2.0.10}$$

It is possible to show that

$$\mathrm{cov}(X, Y) = E[XY] = \iint xy p_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = \rho.$$

Both X, Y have marginal distributions which are standard normal and conditional distributions which are also normal but with a mean that depends on the conditioning argument:

$$X \sim N(0, 1), Y \sim N(0, 1), \qquad Y|X = x \sim N(\rho x, (1-\rho^2)).$$

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

2.1 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 parameter 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 (for example, a price elasticity is not expected to be less than -50). Information can also be derived from prior analyses of other data, including data which is only loosely related to the data set under investigation.

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 2.2 below) 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.

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:

$$p(\theta|D) = \frac{p(D, \theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{2.1.1}$$

$p(\theta|D)$ is called the *posterior* distribution and reflects the combined data and prior information. Equation (2.1.1) is often expressed using the likelihood function. Given D, any function which is proportional to $p(D|\theta)$ is called the *likelihood*, $\ell(\theta)$. The shape of the posterior is determined entirely by the likelihood and prior in the numerator of (2.1.1) and this is often emphasized by rewriting the equation:

$$p(\theta|D) \propto \ell(\theta)p(\theta). \tag{2.1.2}$$

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

TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
WEISBERG · Applied Linear Regression, *Second Edition*
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization
YANG · The Construction Theory of Denumerable Markov Processes
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions: Applications in the Health Sciences
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.

`bayesm` MCMC functions

Name	Description	Model	Priors	Required arguments			Output
				Data	Prior	Mcmc	
`rbprobitGibbs`	Gibbs sampler for binary probit	$z = X\beta + \varepsilon, \varepsilon \sim N(0,1)$ $y = 1$ if $z > 0$	$\beta \sim N(\bar{\beta}, A^{-1})$	X, y		R (no. of MCMC draws)	betadraw
`rhierBinLogit`	Hybrid sampler for hierarchical binary logit – designed for conjoint data. Note: x variables are expressed as a difference between alternatives. See Appendix A.	$y_{hi} = 1$ with $\text{pr} = \dfrac{\exp(x_{hi}'\beta_h)}{1+\exp(x_{hi}'\beta_h)}$ $h = 1, \ldots, H \quad ; i = 1, \ldots, I_h$ H panel units $\beta_h \sim N(\Delta' z_h, V_\beta)$	$V_\beta \sim \text{IW}(\nu, V)$ $\text{vec}(\Delta) \sim N(\text{vec}(\bar{\Delta}), V_\beta \otimes A^{-1})$	lgtdata lgtdata[[h]]$X is X matrix for unit h lgtdata[[h]]$y is dep. var. for unit h		R	betadraw Deltadraw Vbetadraw llike reject
`rhierLinearModel`	Gibbs sampler for linear hierarchical model	$y_i = X_i\beta_i + \varepsilon_i, \varepsilon_i \sim N(0, \tau_i)$ $\beta_i \sim N(\Delta' z_i, V_\beta)$	$\tau_i \sim \nu_\varepsilon \text{ssq}_i / \chi^2_{\nu_\varepsilon}$ $V_\beta \sim \text{IW}(\nu, V)$ $\text{vec}(\Delta) \sim N(\text{vec}(\bar{\Delta}), V_\beta \otimes A^{-1})$	regdata regdata[[i]]$X is X matrix for unit i regdata[[i]]$y is dep. var. for unit i		R	betadraw taudraw Deltadraw Vbetadraw
`rhierMnlRwMixture`	Hybrid sampler for hierarchical logit with mixture of normals prior	$y_i \sim \text{Multinomial}(P_r(X_i, \beta_i))$, $i = 1, \ldots,$ no. of units $\beta_i \sim \Delta' z_i + u_i$ $u_i \sim N(\mu_{\text{ind}_i}, \Sigma_{\text{ind}_i})$ $\text{ind}_i \sim \text{Multinomial}(pvec)$ $pvec$ is ncomp -vector	$\text{pvec} \sim \text{Dirichlet}(a)$ $\text{vec}(\Delta) \sim N(\bar{\delta}, A_\delta^{-1})$ $\mu_j \sim N(\bar{\mu}, \Sigma_j \otimes a_\mu^{-1})$ $\Sigma_j \sim \text{IW}(\nu, V)$	p, lgtdata p is no. of choices lgtdata[[i]]$y is vector of multinomial choices for unit i lgtdata[[i]]$X is X matrix (use createX to make) for ith unit.	ncomp (no. of normal comps)	R	Deltadraw betadraw probdraw compdraw a list of lists of draws of normal comps
`rivGibbs`	Gibbs sampler for linear instrumental variables model	$x_i = z_i'\delta + \varepsilon_1$ $y_i = \beta x_i + w_i'\gamma + \varepsilon_2$ $\varepsilon \sim N(0, \Sigma)$	$\delta \sim N(\mu_\delta, A_\delta^{-1})$ $\text{vec}(\beta, \gamma) \sim N(\mu_{\beta\delta}, A_{\beta\delta}^{-1})$ $\Sigma \sim \text{IW}(\nu, V)$	z, w, x, y z, w are matrices of intstruments, exog. vars x, y are nobs -vectors		R	deltadraw betadraw gammadraw Sigmadraw

`rmnlIndepMetrop`	Independence Metropolis sampler for Multinomial logit model	$y \sim$ Multinomial$(P_r(X, \beta))$	$\beta \sim N(\overline{\beta}, A^{-1})$	p, y, X p is no. of choices X is nobs $*m$ (use `createX`) y is nobs -vector of multinomial indicat.		R	betadraw
`rmnpGibbs`	Gibbs sampler for multinomial probit	$w_i = X_i\beta + \varepsilon, \varepsilon \sim N(0, \Sigma)$ $y_i = j$, if $w_{ij} > \max(w_{i,-j}, 0)$ $j = 1, \ldots, p-1$ else $y_i = p$	$\beta \sim N(\overline{\beta}, A^{-1})$ $\Sigma \sim$ IW(ν, V)	p, y, X p is no. of choices y is nobs -vector of multinomial indicat. X is nobs $*(p-1) \times k$ use createX(DIFF=TRUE) to make X		R	betadraw sigmadraw normalize beta and Sigma draws to identified quantities!
`rmvpGibbs`	Gibbs sampler for the multivariate probit	$w_i = X_i\beta + \varepsilon_i, \varepsilon_i \sim N(0, \Sigma)$ $y_{ij} = 1$ if $w_{ij} > 0$, else 0 $j = 1, \ldots, p$	$\beta \sim N(\overline{\beta}, A^{-1})$ $\Sigma \sim$ IW(ν, V)	p, y, X p is no. of choices y is nobs -vector of multinomial indicat. X is nobs $*p \times k$		R	betadraw sigmadraw normalize to R and divide beta draws by std dev. for ident!
`rnmixGibbs`	Gibbs sampler for mixture of normals	$y_i \sim$ N$(\mu_{\text{ind}_i}, \Sigma_{\text{ind}_i})$ ind$_i \sim$ Multinomial(pvec) pvec is ncomp -vector	pvec $\sim$ Dirichlet(a) $\mu_j \sim$ N$(\overline{\mu}, \Sigma_j \otimes a^{-1})$ $\Sigma_j \sim$ IW(ν, V)	y is $n \times k$ matrix of k-dim. obs	ncomp	R	probdraw zdraw (indicators) compdraw a list of lists of draws of normal comps
`rscaleUsage`	Gibbs sampler for multivariate ordinal data with scale usage heterogeneity. Designed for survey questions on k-point scale	$x_{ij} = d$ if $c_{d-1} < y_{ij} < c_d$ $d = 1, \ldots, k$ $c_d = a + bd + ed^2$ $y_i = \mu + \tau_i\iota + \sigma_i z_i$ $z_i \sim N(0, \Sigma)$	vec$(\tau_i, \log(\sigma_i)) \sim N(\phi, \Lambda)$ $\phi = \begin{pmatrix} 0 \\ \lambda_{22} \end{pmatrix}$ $\mu \sim N(\overline{\mu}, A_\mu^{-1})$ $\Sigma \sim$ IW(ν, V), $\Lambda \sim$ IW(ν_Λ, V_Λ) $e \sim$ unif. on grid	k, x x is nobs $\times$ nvar matrix of responses to nvar survey questions, each one is on k-point scale			Sigmadraw mudraw taudraw sigmadraw Lambdadraw edraw use defaults!
`runiregGibbs`	Gibbs sampler for univariate regression model	$y = X\beta + \varepsilon, \varepsilon \sim$ N$(0, \sigma^2 I_{nobs})$	$\beta \sim N(\overline{\beta}, A^{-1})$ $\sigma^2 \sim \nu\text{ssq}/\chi^2_\nu$	y, X		R	betadraw sigmasqdraw

Key `bayesm` utilities

Name	Description	Arguments	Output	Notes
`condMom`	Moments of *i*th element of multivariate normal given others	*x* vector of conditioning values (*i*th not used) mu mean sigi inverse of covariance matrix *i* index of element to compute cond dist for	list(cmean,cvar) conditional mean and variance	
`createX`	Creates *X* matrix for MNL and probit routines	*p* number of choice alternatives na number of alternative specific vars in Xa nd number of non-alternative specific vars (demos) Xa n*p×na matrix of alternative specific vars Xd n × nd matrix of non-alternative specific vars INT logical flag for intercepts DIFF logical flag to diff. with respect to base alternative base number of base alternative	*X* matrix	Use with rmnlIndepMetrop, rhierMnlRwMixture, rmnpGibbs, rmvpGibbs, llmnl, llmnp
`ghkvec`	Computes GHK approximation to integral of normal over half plane defined by vector of truncation points	*L* lower Cholesky root of Σ matrix trunpt vector of truncation points (see notes) above vector of indicators for above(1) or below(0) *r* number of draws in GHK	approx. to integral. vector of length= length(trunpt) /length(above)	Allows for same density and same side of axes but different truncation points to produce a vector
`llmnl`	Log-likelihood for MNL logit model	beta is *k*-vector of coefficients *y* is *n*-vector of multinomial outcomes from p alternatives *X* is n*p xk design matrix (use createX)	value of log-likelihood	
`llmnp`	Log-likelihood for MNP model	beta is *k*-vector of coefficients Sigma is covariance matrix of errors *X* $n^*(p-1) \times k$ array (use createX with DIFF on) *y* *n*-vector of multinomial outcomes (*p* alternative) *r* is number of draws to use in GHK	value of log-likelihood	