Chapter 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 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, Gelman et al, and Berger 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 are models 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 non-uniform 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 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 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 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, 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.

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

If we observe more than one observation per consumer, it is natural to consider a model which accommodates differences between consumers. That is, we have some information about each consumers 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.

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

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 (1.2.2). 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]

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

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.

[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.

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.

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

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 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.

1.3 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 which 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 non-trivial 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 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 two fold. 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 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 parameters spaces arise either due to large numbers of "units," or to 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 non-parametric 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 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 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 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 exogeneous variables. In marketing applications, the marketing mix variables and sales are joint 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 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 is 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 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 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 80s and early 90s. 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. 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 don't claim to have come any where 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 500 MB is now relatively standard for Windows machines, we do not think this is unreasonable as much as it might raise eyebrows of 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 x 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: 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 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 B for more information on *bayesm* and appendix A for an introduction to R). The book's website, xxxxx, 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.

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, 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. Greg would like to thank Vijay Bhargava for his encouragement in both personal and professional life. Finally, our wives, Tricia, Nancy, and Laurie, have put up with a certain amount of crankiness associated with our frustration at the progress of this research.