Bayesian Statistics and Marketing

Peter E. Rossi • Greg M. Allenby
Graduate School of Business, University of Chicago, 1101 E. 58th Street, Chicago, Illinois 60637
Fisher College of Business, Ohio State University, 2100 Neil Avenue, Columbus, Ohio 43210
peter.rossi@gsb.uchicago.edu • allenby.1@osu.edu

Bayesian methods have become widespread in marketing literature. We review the essence of the Bayesian approach and explain why it is particularly useful for marketing problems. While the appeal of the Bayesian approach has long been noted by researchers, recent developments in computational methods and expanded availability of detailed marketplace data has fueled the growth in application of Bayesian methods in marketing. We emphasize the modularity and flexibility of modern Bayesian approaches. The usefulness of Bayesian methods in situations in which there is limited information about a large number of units or where the information comes from different sources is noted. We include an extensive discussion of open issues and directions for future research.

(*Bayesian Statistics; Decision Theory; Marketing Models; Critical Review*)

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. Bayesian methods are particularly appropriate to the decision orientation of marketing problems. 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. In this paper, we will outline the basic advantages of the Bayesian approach, explain how hierarchical Bayes models are ideally suited to many marketing data sets and decisions, and outline the nature of the computational revolution. Throughout, we will emphasize the importance of a decision orientation that we believe is an important aspect of marketing as a field.

Until the mid-1980s, Bayesian methods appeared to be impractical because the class of models for which the posterior could be computed were 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 as discussed below.

While Bayesian methods have risen to prominence in many fields, this review will emphasize a perspective on the use of Bayes methods that stems from a basic marketing paradigm. Fundamental to this perspective is the notion that customers are different in their preferences for products and that firms must explicitly take this into account in determining optimal marketing actions. It is useful, therefore, to view statistical analysis as comprised of three components:

1. within-unit behavior (the conditional likelihood);
2. across-unit behavior (the distribution of heterogeneity);
3. action (the solution to a decision problem involving a loss function).

0732-2399/03/2203/0304
1526-548X electronic ISSN

We will see how the Bayesian approach provides a unified treatment of all three components.

We will follow these three steps as the outline of the paper, and conclude the paper with a discussion of open issues and directions for future research. We have also included Annotated Citations of Bayesian Applications in Marketing in Appendix 1, which contains a list of published or accepted papers of the last ten years that tackle marketing problems using Bayesian methods. The annotations provide a brief description of the paper and how it relates to the topics discussed in this paper.

2. Bayesian Essentials

In this section, we introduce our notation for the Bayesian paradigm, and comment on the important distinctions between classical and Bayesian approaches. We feel that these distinctions are underappreciated by researchers in marketing. We do not attempt to provide a primer for Bayesian inference. For those interested in an introduction to Bayesian inference and modern Bayesian computing methods, there are many excellent texts, including Bernardo and Smith (1994), Gelman et al. (1995), Robert and Casella (1999), and Liu (2001).

All Bayesian analysis starts with the specification of the data-generating mechanism or the distribution of the data y, given the unobservable parameters θ, $p(y \mid \theta)$. Viewed as a function of the parameters, this distribution is sometimes called the likelihood function, $l(\theta) = p(y \mid \theta)$. The Bayesian, therefore, subscribes to the likelihood principle that states that the likelihood function contains all relevant information regarding the model parameters. In addition, a probability distribution representing prior beliefs about θ is required, $p(\theta)$. Bayes theorem provides the updating mechanism for how prior beliefs are translated into posterior (or after the data) beliefs.

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \propto p(y \mid \theta)p(\theta).$$

$p(\theta \mid y)$ is called the posterior distribution and reflects both the prior beliefs, as well as sample information. We note, immediately, that the posterior is a conditional distribution that conditions on the data. This provides a marked contrast to the sampling theoretic view in which we consider the data random, and we investigate the behavior of test statistics or estimators over imaginary samples from $p(y \mid \theta)$. The Bayesian would regard the sampling distribution as irrelevant to the problem of inference because it considers events (y) that have not occurred. Inference is the problem of making statements about the unobservables *conditional* on the data.

Since the posterior distribution can be a high-dimensional object, investigators typically summarize the posterior in terms of some lower dimensional summary statistics. Typically, the posterior mean $E[\theta] = \int \theta\, p(\theta \mid y)\, d\theta$ is used as an estimator and the posterior standard deviation is used as a measure of precision. Both of these quantities are the integrals of specific functions of the parameter vector, $E_{\theta \mid y}[h(\theta)]$. Other important examples include: (i) Aspects of the marginal distribution of one element or a subset of the θ vector; (ii) posterior probabilities of intervals or regions of the parameter space (such as the posterior probability that a price coefficient is negative); and (iii) predictive distributions of the data, $p(y_f \mid y) = \int p(y_f \mid \theta) p(\theta \mid y)\, d\theta$. Thus, the Bayesian investigator is faced with the problem of computing a multidimensional integral of the posterior distribution. Methods for computing these integrals are at the core of the recent revolution in computing for Bayesian statistics.

The Bayesian framework is compelling in the sense that it provides a unified approach to modeling, incorporation of prior information, and inference. Inference here refers to making a posteriori statements about all unobservables including both parameters and, as yet unrealized, data (prediction). Bayesian inference adheres to the likelihood principle and is conducted using formal rules of probability theory. This means that, under mild conditions, Bayes estimators are consistent, asymptotically efficient, and admissible. As a practical matter, Bayesian inference is free from the use of asymptotic approximations and delivers exact, finite sample inference. This is particularly important in nonlinear models and models with discrete data. The intuition developed for regression models of the sample size required for asymptotic sampling theory to be accurate does not carry over well to many of the models used with marketing data. In particular, choice models may require extremely large (as much as 1,000 observations per parameter) samples to insure the adequacy of asymptotic approximations (cf. McCulloch and Rossi 1994).

In general, Bayesian methods provide a better approximation to the level of uncertainty or, conversely, the amount of information provided by the model and the data than other approaches. For example, consider two-step procedures in which a subset of parameters are estimated in the first stage, then the second stage estimates the remaining parameters, conditional on the first subset. Parameter uncertainty is difficult to account for in multistage analyses. Lenk and DeSarbo (2000) provide an example of how a full Bayesian procedure outperforms an approximate two-step procedure for clustering problems. Parameter uncertainty and model uncertainty are particularly important considerations in optimal decision theory. Optimal decision making should take into account uncertainty to avoid the problem of "overconfidence" (see Montgomery and Bradlow 1999). Bayesian decision theory provides a unified approach to inference, model choice, and uncertainty as discussed in §6 of this paper

The advantages of Bayesian inference are not obtained without a cost, however. The Bayesian approach is likelihood-based and requires a prior. Some have criticized Bayesian methods as relying on "subjective" prior information. It is important to note that the basis of prior information can also be "objective" or data-based. In addition, all modeling assumptions are a form of prior information. The advantage of the Bayesian approach is that all prior assumptions are explicitly stated. Adherence to the principles of scientific inquiry does not rule out the use of subjective information but, rather, the specification of explicit and replicable procedures. It should be noted that in the practical domain of marketing, methods that make full use of prior information are required for reliable inference because information about unknown quantities is hard to come by. Prior information from experts (Sandor and Wedel 2001), theories (Montgomery and Rossi 1999), or other datasets (Lenk and Rao 1990, Putler et al. 1996, Kamakura and Wedel 1997, Wedel and Pieters 2000, Ansari et al. 2000a, Ter Hofstede et al. 2002) is important to marketing problems. Classical inference procedures are silent on how to incorporate information from sources other than the data.

Some contend that specification of the likelihood function is another drawback of the Bayesian approach. For some models, evaluation of the likelihood can be computationally demanding. In other situations, the investigator may be concerned with model specification error induced by specifying an inappropriate likelihood. Recent developments in statistical computing have opened up the possibility of analyzing likelihood functions once thought to be computationally intractable. Regarding prior and likelihood specification, we recommend that the investigator perform sensitivity analysis.

3. MCMC Simulation Methods

The general computational problem facing Bayesians is the computation of various integrals of functions with respect to the posterior distribution. Since these integrals can be written as the posterior expectation of a function of the parameters, simulation methods seem natural candidates for approximation. For example, if we could make i.i.d. draws from the posterior we could simply approximate the integrals by the sample mean

$$I = E_{\theta|y}[h(\theta)] = \int h(\theta)p(\theta \mid y)\,d\theta$$
$$\hat{I} = 1/R\sum_{r=1}^{R} h(\theta_r).$$

If draws from the posterior are available at low computational cost, we could simply use a very large sample to approximate I to any desired degree of accuracy. However, the general problem of drawing from an arbitrary multivariate distribution is extremely difficult and there is no computationally feasible general method.

Instead of using i.i.d. draws, another approach could be to construct a Markov chain with the posterior as its stationary or equilibrium distribution. In practice, this means specifying a transition density that produces a sequence of θ draws. θ_r is a draw from $p(\theta_r \mid \theta_{r-1})$ given θ_0. If $p(\theta \mid y)$ is the stationary distribution of this Markov chain, then we can simply iterate the chain long enough to dissipate the effects of the initial condition and then save these draws to evaluate $\hat{I}$. While these draws are no longer i.i.d. (they will exhibit some form of autocorrelation in most cases), laws of large numbers still apply and we can approximate I to any desired degree of accuracy. The use of a Markov chain to develop a simulation-based estimate of I has been termed a MCMC method (see Robert and Casella 1999 for a comprehensive discussion of these methods, and Chib 2003 for an excellent overview). The usefulness of the MCMC idea depends on three criteria:

(i) the ability to construct chains for arbitrary posterior distributions;

(ii) the ability of the chain to quickly converge to the equilibrium distribution and not to exhibit highly autocorrelated or near nonstationary behavior;

(iii) ease of drawing from the transition density of the chain.

The class of Metropolis-Hastings algorithms provide a set of methods for constructing Markov chains. Tierney (1994) shows that under very mild conditions (mostly that the posterior density is positive everywhere in the parameter space), the Metropolis-Hastings style methods will converge at a geometric rate to the unique equilibrium distribution that will be the posterior. One particularly useful member of the Metropolis-Hastings class is the so-called Gibbs sampler. The Gibbs sampler is dependent on the ability to draw from various conditional distributions of the joint posterior. Partition the θ vector into K subvectors, $\theta' = (\theta'_1, \ldots, \theta'_K)$, and consider the conditional distributions $p_k(\theta_k \mid \theta_{-k}, y)$, where $-k$ refers to the elements of θ, other than the kth element. If it is possible to draw from these conditional distributions, then the Markov chain that can be constructed by cycling through these conditional distributions has the posterior as its invariant distribution. That is, to draw $\theta_r \mid \theta_{r-1}$, we cycle through each of the K conditionals

$$\begin{gathered} p_1\big(\theta_{r,1} \mid \theta_{r-1,2}, \theta_{r-1,3}, \ldots, \theta_{r-1,K}, y\big) \\ p_2\big(\theta_{r,2} \mid \theta_{r,1}, \theta_{r-1,3}, \ldots, \theta_{r-1,K}, y\big) \\ \vdots \\ p_K\big(\theta_{r,K} \mid \theta_{r,1}, \theta_{r,2}, \ldots, \theta_{r,K-1}, y\big). \end{gathered}$$

The sequence of draws converges in distribution to the joint posterior distribution of the model parameters, $p(\theta_1, \ldots, \theta_K \mid y)$. In addition, draws from the posterior distribution for any one parameter,

$$p(\theta_i \mid y) = \int p(\theta_1, \ldots, \theta_K \mid y)\, d\theta_{-i}$$

is obtained by simply discarding the draws of parameters not of interest. For example, the posterior mean of $p(\theta_i \mid y)$ can be estimated from the sample mean of the θ_i draws.

Many models can be expressed such in a way that these various conditionals are available in closed-form and from well-known distributions that are easy to sample from. For example, single and groups of linear regressions fall into this class. In addition, data augmentation allows standard probit models to be sampled using the Gibbs sampler. Today most work is being done with models that no longer have a simple Gibbs sampler. However, the modular or conditional setup of the Gibbs sampler is frequently exploited. Typically, some sort of hybrid approach is used, in which Gibbs-style draws are combined with "Metropolis"-style draws.

It is possible to define a Metropolis-Hastings style MCMC algorithm for many models, including highly nonlinear models or models defined in high dimensions. These algorithms, however, must be investigated closely to insure that they navigate freely through the parameter space and reach the regions of high posterior probability. It is possible to stop a slowly navigating chain and conclude that the posterior is very tight when, in fact, the algorithm is moving too slowly. We recommend that investigators simulate data to investigate performance of the MCMC algorithm. Independence Metropolis or random walk Metropolis algorithms must have properly chosen candidate sampling distributions in order to function well in high-dimensional parameter spaces. We recommend that investigators check these methods against the slower, but more reliable, one-by-one Griddy Gibbs methods.

In summary, MCMC methods are now available to handle inference in a wide class of models. MCMC methods are particularly well-suited to models which are built from a hierarchy of conditional distributions. One very important example is random coefficient models, discussed below. Perhaps, more importantly, the modularity of the hierarchical modeling approaches, that dovetails so well with MCMC methods, has enlarged the class of priors and likelihoods available for use in marketing applications.

4. Within-Unit Analysis: Likelihoods and Marketing Data

Marketing is concerned with understanding and reacting to the behavior of individual consumers. Decisions are ultimately made at a disaggregate level, although for some types of decisions (e.g., setting store prices) an aggregate-level analysis is acceptable. We note that it is always possible to derive aggregate predictions of actions by integrating over the distribution of heterogeneity. Our discussion, therefore, focuses on data and models assuming that the unit of analysis is an individual respondent, consumer, or household.

Marketing data is sparse at the individual-unit level. In scanner panels of household purchases, for example, it is rare to have more than 20 observations per household in most product categories. Each observation is a vector response corresponding to the quantity purchased of a particular offering. The most frequent response value is zero, indicating no purchase of the offering, and the second most frequent response is one, indicating that one unit of the good is purchased. Responses also take on integer values in surveys where respondents are asked to choose between discrete alternatives, to rank order objects (Bradlow and Fader 2001), and to provide responses on five- and seven-point scales. Marketing data are typically very lumpy, and are not well-suited to standard distributional assumptions (e.g., normal, gamma, Poisson).

Latent variable models are often used to explain marketing data. A latent variable model typically assumes that there exists an unobserved continuous variable and a censoring mechanism that gives rise to the discrete outcome. In an economic model of choice between near-perfect substitutes, for example, consumers are assumed to select the offering with greatest value, measured as the ratio of marginal utility to price. In the analysis of survey response data, it is sometimes convenient to assume that responses on a fixed-point (e.g., five- and seven-point) scale are a censored realization of a latent, continuous variable. Finally, in the analysis of multiple response data (i.e., pick any of J data), each element of the vector of multivariate binomial responses can be thought of as being equal to one if a latent variable surpasses a threshold, and equal to zero if the latent variable is less than the threshold value.

The advantage of a latent variable approach to modeling marketing data is that it provides a flexible approach to specifying the likelihood function that is consistent with the observed, lumpy data (Rossi et al. 2001, Marshall and Bradlow 2002). Models for the latent utility can be continuous even though the range of the observed dependent variable is discrete. Many useful models can be constructed starting from an underlying multivariate normal regression model

$$z = X\beta + \varepsilon \qquad \varepsilon \sim N(0, \Sigma).$$

Here, z is a $m \times 1$ vector, which is multivariate normal conditional on x. The latent vector, z, is censored via some function which is not a function of the model parameters, (β, Σ). Examples include:

Tobit Model: $m = 1$, $y = 0$, $z < 0$, $y = z$, $z \geq 0$

Ordered Probit: $m = 1$, $y = r$, $c_{r-1} \leq z < c_r$, $r = 1, \ldots, R$, $c_0 = -\infty$, $c_R = \infty$

Multinomial Probit (MNP): $y = j$, $z_j = \max(z_1, \ldots, z_m)$

Multivariate Probit: $y_j = 1$, $z_j > 0$; else $y_j = 0$

These four examples illustrate the flexibility of the latent framework. The Tobit model produces a discrete-continuous distribution for y given x that has a lump of probability at zero (the no-purchase option, for example). The ordered probit can be applied to ratings data in which the respondent provides ratings on a ratings scale. The MNP probit model is a very flexible general model that accommodates situations in which choices are made from a set of m alternatives. Finally, the multivariate probit model can be used in situations such as the pick j from J alternatives or where binary choice is made in different time periods or categories of products.

Latent variable models can often be given an economic interpretation as a random utility model. Consider, for example, the MNP model. If consumers have linear utility and can only choose one alternative, the utility-maximizing choice is the choice for which the ratio of marginal utility to price is the highest;

$$y = j; \qquad \text{if } U_j/p_j = \max\{U_i/p_i\},$$

where U_i is the marginal utility of choice i. In the random utility model, marginal utilities are not fully observable. We only observe various attributes of the choice that are represented in the x vector. If $\ln(U_i) = V_i + \varepsilon_i$ and $\varepsilon_i \sim N(0, \Sigma)$, then the model becomes

$$y = j; \qquad \text{if } V_j - \ln(p_j) + \varepsilon_j = \max_i\{V_i - \ln(p_i) + \varepsilon_i\}.$$

This is a special case of the MNP model. The error terms have the interpretation as the unobservable factors influencing marginal utility. The random utility approach can be applied to any demand model (see Blattberg and George 1991, Arora et al. 1998, Manchanda et al. 1999, Bradlow and Rao 2000, Leichty et al. 2001). If we specify a utility function, the random component of marginal utility will induce a distribution on the quantity demanded via the first order conditions for utility maximization. This will create a likelihood for the data. If the indifference curves of the specified utility function intersect the axes of the positive orthant with nonzero slope, then there is the potential for corner solutions in which some of the components of the demand vector will be zero. These corners will create a mixed discrete-continuous distribution of demand (Kim et al. 2002).

Models involving multivariate latent variables (such as the MNP and multivariate probit models) have a likelihood function that can be computationally challenging to evaluate. For example, consider the MNP model. If alternative j is chosen from m alternatives, this reveals that we are in a certain region of the error space (actually a cone). Thus, the multinomial probabilities required for evaluation of the MNP likelihood involve integrals over a region of the error space

$$\Pr(i \mid \beta, \Sigma) = \int_R \phi(z \mid \mu = X\beta, \Sigma)\, dz.$$

The choice probabilities involve integrals of a multivariate normal density over cones, and these integrals pose a potentially severe computational problem. Classical econometricians have focused on methods for approximating these integrals. The state of the art in this area is the so-called GHK algorithm (Keane 1993). The GHK algorithm uses importance sampling to approximate these probabilities. The current classical practice involves using simulation methods to approximate the likelihood (Huber and Train 2001) and then uses standard maximum likelihood procedures, ignoring the simulation error (this is often called the simulated maximum likelihood approach).

4.1. Data Augmentation

Direct evaluation of the censored normal likelihood can be avoided in a Bayesian approach if the parameters are *augmented* with a vector of latent variables, z (see Tanner and Wong 1987). To a Bayesian, all unobservable quantities can be considered the object of inference regardless of whether they are called parameters or latent variables. Technically, the number of latent variables can be the same as the number of observations, so large sample inference based on standard asymptotics does not apply. The posterior we now require is the joint distribution of the unobservable latent vector (z), and the parameter vector (θ), given the data (y). To reduce notational burden, we will only consider the case of one observation. The joint posterior of z and θ is now the object of inference. The posterior of θ is a marginal of this joint posterior.

$$p(\theta \mid y) = \int p(z, \theta \mid y)\, dz.$$

As it turns out, we can exploit the latent structure of the model to construct a "Gibbs-style" Markov chain that can sample from the joint posterior of the latents and the parameters. We can then simply marginalize on the data by discarding the draws of z. That is, we draw iteratively from the two conditional distributions:

$$p(z \mid y, \theta) \quad \text{and} \quad p(\theta \mid z, y).$$

The draw of the latent z given y is a draw from a truncated normal distribution where the truncation depends on the model. In the MNP case, z is truncated to a m-dimensional cone. Given the latent vector z, inference proceeds as would standard Bayesian analysis of the underlying latent multivariate regression model. For the linear multivariate regression model, exact analytic results are available for the posterior of (β, Σ). Draws from the truncated multivariate normal can easily be accomplished via one-by-one draws from a series of univariate truncated normal distributions (see McCulloch and Rossi 1994, Allenby et al. 1995). This amounts to defining a subchain to draw the truncated normal vector. What is important to note is that by augmenting with the latent variable, we have avoided evaluation of any choice probabilities or other integrals of the multivariate normal. The cost of computational simplification is an enlargement of the state space for the Markov chain. In general, this will cause the data-augmented MCMC method to converge more slowly and exhibit higher autocorrelation than the non-data-augmented sampler. In the case of the MNP model, Nobile (1998) has indicated that, under certain conditions, the standard augmented Gibbs sampler can exhibit very high autocorrelation and proposes an improved chain.

Thus, data augmentation provides a clever way of avoiding evaluation of various multivariate integrals at the possible expense of introducing high autocorrelation in to the MCMC method. Our experience, however, has shown that the basic MNP Gibbs sampler works well and can handle problems for which the method of simulated maximum likelihood grinds to a halt.

4.2. Identification

The latent variable formulation provides a natural mechanism for understanding the identification problem in these models. Identification problems stem from the fact that various transformations of the latent variables leave the observed censored outcome variable unchanged. For example, recall that in the MNP model the choice is made with the highest latent value. There are two transformations that leave the index of the maximum unchanged—location and scale shifts (see McCulloch and Rossi 1999, for further details). Identification can be achieved either by imposing exact restrictions on the model parameters, or by employing informative priors on the full parameter space and marginalizing on the identified parameters.

In many classical and Bayesian approaches, the approach to this scaling problem is to fix one of the elements of the covariance matrix (typically, the (1, 1) element) to one. For Bayesian methods, the restriction of the covariance matrix makes it difficult to use standard conjugate priors such as the Wishart prior. McCulloch et al. (2000) show how to construct practical priors on the appropriate space of matrices.

However, Bayesians are not limited to exact restrictions as a way of solving various identification problems. Use of a proper prior distribution ensures that the posterior is proper, even if the likelihood is not identified. In a Bayesian analysis, the issue of statistical identification shifts from an identified—not identified dichotomy, to an issue of the degree of identification and to subspaces of the posterior distribution that are well identified. For example, we can use a proper but diffuse prior in the unidentified parameter space, and simply marginalize or project down on the space of identified parameters. The only added cost of this procedure is making sure that the induced prior on the identified quantities is sufficiently diffuse to be usable in those situations in which we want our inferences driven primarily by the data.

An even more striking example of the usefulness of this idea of navigating in the full, unidentified space can be found in the multivariate probit model. Here the identified parameters consist only of the correlation matrix of the latent variables because separate scaling constants can be used for each element. Until recently (see Barnard et al. 2000) convenient priors for correlation matrices have not be available. Standard MCMC methods, such as Metropolis-Hastings, are difficult to adapt to the highly restricted space of valid correlation matrices (Manchanda et al. 1999). In other words, it is hard to draw candidate correlation matrices. As Edwards and Allenby (2002) illustrate, all of this can be avoided by navigating in the unidentified space and projecting down to the space of correlation matrices (see also DeSarbo et al. 1999). These algorithms are fast and reliable.

We have seen that disaggregate marketing data is often lumpy, containing discrete mass points of probability. A natural framework for building models with discrete aspects is to use an underlying continuous latent variable, coupled with some sort of censoring mechanism. Not only are latent variables useful for generating models but also the new MCMC Bayesian inference methods nicely exploit the latent structure. Finally, identification problems that are common in latent variable models can be handled with great flexibility in the Bayesian approach.

5. Across-Unit Analysis: Incorporating Heterogeneity via Hierarchical Models

The explosion in demand data available to marketers comes from the increased availability of disaggregate data. Scanner data at the store and household level is now commonplace. In the pharmaceutical industry, physician-level prescription data is now commonplace. This raises both modeling challenges, as well as major opportunities for improved profitability through decentralized marketing decisions that exploit heterogeneity. This new data comes in panel structure in which N, the number of units is large relative to T, the length of the panel. Thus, we may have a large amount of data obtained by observing a large number of decision units. For a variety of reasons, it is unlikely that we will ever have a very large amount of information about any one decision unit. In this situation, it is useful to have a model that pools information among the units. A flexible random effects model, combined with Bayesian inference methods, can produce accurate estimates at both the aggregate and individual decision unit level.

5.1. Heterogeneity and Priors

A useful general structure for disaggregate data is a panel structure in which the units are regarded as independent, conditional on unit-level parameters. Given a joint prior on the collection of unit-level parameters, the posterior distribution can be written as follows:

$$p(\theta_1, \ldots, \theta_N \mid y_1, \ldots, y_N) \propto \left[\prod_i p(y_i \mid \theta_i)\right] \times p(\theta_1, \ldots, \theta_N \mid \tau).$$

The term in brackets is the conditional likelihood and the rightmost term is the joint prior with hyperparameter, τ. In many instances, the amount of information available for many of the units is small. This means that the specification of the functional form and hyperparameter for the prior may be important in determining the inferences made for any one unit. A good example of this can be found in choice data sets in which consumers are observed to be choosing from a set of products. Many consumers ("units") do not choose all of the alternatives available during the course of observation. In this situation, most standard choice models do not have a bounded maximum likelihood estimate (the likelihood has an asymptote in a certain direction in the parameter space). In this situation, the prior is, in large part, determining the inferences made for these consumers.

Assessment of the joint prior for $(\theta_1, \ldots, \theta_N)$ is difficult, due to the high dimension of the parameter space and, therefore, some sort of simplification of the form of the prior is required. One frequently employed simplification is to assume that, conditional on the hyperparameter, $(\theta_1, \ldots, \theta_N)$ are a priori independent.

$$p(\theta_1, \ldots, \theta_N \mid y_1, \ldots, y_N) \propto \prod_i p(y_i \mid \theta_i) p(\theta_i \mid \tau).$$

This means that inference for each unit can be conducted independently of all other units *conditional* on τ. This is the Bayesian analogue of fixed-effects approaches in classical statistics.

The specification of the conditionally independent prior can be very important, due to the scarcity of data for many of the units. Both the form of the prior and the values of the hyperparameters are important and can have pronounced effects on the unit-level inferences. For example, it is common to specify a normal prior, $\theta_i \sim N(\bar{\theta}, V_\theta)$. The normal form of this prior means that influence of the likelihood for each unit may be attenuated for likelihoods centered far away from the prior. That is, the thin tails of the normal distribution diminish the influence of outlying observations. In this sense, the specification of a normal form for the prior, whatever the values of the hyperparameters, is far from innocuous.

Assessment of the prior hyperparameters can also be challenging in any applied situation. For the case of the normal prior, some relatively diffuse prior may be a reasonable default choice. Rossi and Allenby (1993) use a prior, based on a scaled version of the pooled model information matrix. The prior covariance is scaled back to represent the expected information in one observation to insure a relatively diffuse prior. Use of this sort of normal prior will induce a phenomenon of "shrinkage" in which the Bayes estimates (posterior means) $\{\tilde{\theta}_i = E[\theta_i \mid \text{data}_i, \text{prior}]\}$ will be clustered more closely to the prior mean than the unit-level maximum likelihood estimates $\{\hat{\theta}_i\}$. For diffuse prior settings, the normal form of the prior will be responsible for the shrinkage effects. In particular, outliers will be "shrunk" dramatically toward the prior mean. For many applications, this is a very desirable feature of the normal form prior. We will "shrink" the outliers in toward the rest of the parameter estimates and leave the rest pretty much alone.

5.2. Hierarchical Models

In general, however, it may be desirable to have the amount of shrinkage induced by the priors driven by information in the data. That is, we should "adapt" the level of shrinkage to the information in the data regarding the dispersion in $\{\theta_i\}$. If, for example, we observe that the $\{\theta_i\}$ are tightly distributed about some location or that there is very little information in each unit-level likelihood, then we might want to increase the tightness of the prior so that the shrinkage effects are larger. This feature of "adaptive shrinkage" was the original motivation for work by Efron and Morris (1975) and others on empirical Bayes approaches in which prior parameters were estimated. These empirical Bayes approaches are an approximation to a full Bayes approach in which we specify a second-stage prior on the hyperparameters of the conditional independent prior. This specification is called a hierarchical Bayes model and consists of the unit-level likelihood and two stages of priors.

Likelihood:	$p(y_i \mid \theta_i)$
First-stage prior:	$p(\theta_i \mid \tau)$
Second-stage prior:	$p(\tau \mid h)$.

The joint posterior for the hierarchical model is given by

$$p(\theta_1, \ldots, \theta_m, \tau \mid y_1, \ldots, y_m, h) \propto \left[\prod_i p(y_i \mid \theta_i) p(\theta_i \mid \tau)\right] \times p(\tau \mid h).$$

In the hierarchical model, the prior induced on the unit-level parameters is not an independent prior. The unit-level parameters are conditionally, but not unconditionally, a priori independent.

$$p(\theta_1, \ldots, \theta_m \mid h) = \int \prod_i p(\theta_i \mid \tau) p(\tau \mid h)\, d\tau.$$

If, for example, the second-stage prior on τ is very diffuse, the marginal priors on the unit-level parameters, θ_i, will be highly dependent, as each parameter has a large common component.

The hierarchical model specifies that both prior and sample information will be used to make inferences about the common parameter, τ. For example, in normal prior, $\theta_i \sim N(\bar{\theta}, V_\theta)$, the common parameters provide the location and the spread of the distribution of θ_i. Thus, the posterior for the θ_i will reflect a level of shrinkage inferred from the data. It is important to remember, however, that the normal functional form will induce a great deal of shrinkage for outlying units, even if the posterior of V_θ is centered on large values.

5.3. Inference for Hierarchical Models

Hierarchical models for panel data structures are ideally suited for MCMC methods. In particular, a "Gibbs"-style Markov chain can often be constructed by considering the basic two sets of conditionals:

(1) $\theta_i \mid \tau, y_i$

and

(2) $\tau \mid \{\theta_i\}$

The first set of conditionals exploits the fact that the θ_i are conditionally independent. The second set exploits the fact that $\{\theta_i\}$ are sufficient for τ. That is, once the $\{\theta_i\}$ are drawn from (1), these serve as "data" to the inferences regarding τ. If, for example, the first-stage prior is normal, then standard natural conjugate priors can be used, and all draws can be done one-for-one and in logical blocks. This normal prior model is also the building block for other more complicated priors. The normal model is given by

$$\theta_i \sim N(\bar{\theta}, V_\theta)$$

$$\bar{\theta} \sim N(\bar{\bar{\theta}}, A^{-1})$$

$$V_\theta^{-1} \sim W(\upsilon, V).$$

In the normal model, the $\{\theta_i\}$ drawn from (1) are treated as a multivariate normal sample and standard conditionally conjugate priors are used. It is worth noting that in many applications the second-stage priors are set to be very diffuse ($A^{-1} = 100I$ or larger) and the Wishart is set to have expectation I with very small degrees of freedom such as $\dim(\theta) + 3$. As we often have a larger number of units in the analysis, the data seems to overwhelm these priors and we learn a great deal about τ, or in the case of the normal prior, $(\bar{\theta}, V_\theta)$.

In classical approaches to these models, the first-stage prior is called a random effects model and is considered part of the likelihood. The random effects model is used to average the conditional likelihood to produce an unconditional likelihood which is a function of the common parameters alone.

$$l(\tau) = \prod_i \int p(y_i \mid \theta_i) p(\theta_i \mid \tau)\, d\theta_i.$$

In the classic econometric literature, much is made of the distinction between random coefficient models and fixed effect models. Fixed effect models are considered "nonparametric" in the sense that there is no specified distribution for the θ_i parameters. Random coefficient models are often consider more efficient, but subject to specification error in the assumed random effects distribution, $p(\theta_i \mid \tau)$. In a Bayesian treatment, we see that the distinction between these two approaches is in the formulation of the joint prior on $\{\theta_1, \dots, \theta_m\}$.

5.4. Heterogeneity Distributions

Much of the work in both marketing and in the general statistics literature has used the normal prior for the first stage of the hierarchical model. The normal prior offers a great deal of flexibility and fits conveniently with large Bayesian regression/multivariate analysis literature. The standard normal model can easily handle analysis of many units (Steenburgh et al. 2002), and can be extended to include observable determinants of heterogeneity (see Allenby and Ginter 1995, Rossi et al. 1996, Talukdar et al. 2002). This can be done by introducing a multivariate regression in the observables into the mean function

$$\theta = Bz + u$$
$$u \sim N(0, V_\theta).$$

Here, z is a vector of explanatory variables that are meant to explain across-unit differences. Typically, we might postulate that various demographic or market characteristics might explain differences in intercepts (brand preference) or slopes (marketing mix sensitivities). In linear models, these normal prior specifications amount to specifying a set of interactions between the explanatory variables in the model explaining y (see McCulloch and Rossi 1994, for further discussion of this point).

While the normal model is flexible, there are several drawbacks for marketing applications. As discussed above, the thin tails of the normal model tend to shrink outlying units greatly toward the center of the data. While this may be desirable in many applications, it is a drawback in discovering new structure in the data. For example, if the distribution of the unit-level parameters is bimodal (something to be expected in models with brand intercepts), then a normal first-stage prior may shrink the unit-level estimates to such a degree as to mask the multimodality (see below for further discussions of diagnostics). Fortunately, the normal model provides a building block for a mixture of normals extension of the first-stage prior. The mixture of normals model can be written

$$\begin{aligned} p(\theta \mid \bar{\theta}_1, \dots, \bar{\theta}_K, V_1, \dots, V_K) \\ = r_1 \varphi(\theta \mid \bar{\theta}_1, V_1) + \cdots + r_K \varphi(\theta \mid \bar{\theta}_K, V_K); \end{aligned}$$
$$\sum r_k = 1.$$

It is well-known that the mixture of normals model provides a great deal of flexibility and that with enough components, virtually any multivariate density can be approximated. In particular, multiple modes are possible. Fatter tails than the normal can also be accommodated by mixing in normal components with large variance.

The mixture of normals model can be viewed as a generalization of the popular finite mixture model. The finite mixture model views the prior as a discrete distribution with a set of mass points. This approach has been very popular in marketing, due to the interpretation of each mixture point as representing a "segment" and to the ease of estimation. In addition, the finite mixture approach can be given the interpretation of a nonparametric method as in Heckman and Singer (1982). Critics of the finite mixture approach have pointed to the implausibility of the existence of a small number of homogeneous segments, as well as the fact that the finite mixture approach does not allow for extreme units whose parameters lie outside the convex hull of the support points. The mixture of normals approach avoids the drawbacks of the finite mixture model, while incorporating many of the more desirable features.

The MCMC algorithm for the normal heterogeneity model can easily be extended to handle the mixture of normals model by appending indicator variables for the mixture component to the state space. Conditional on the indicator variables, the draws of the normal component parameters are standard conjugate draws given the classification of the observations into one of the K components. The indicator variables, conditional on all other parameters, have a multinomial distribution with probabilities proportional to the number of units assigned to the component and the likelihood that the unit's parameters are from the component distribution.

In mixture of components models, there is a generic identification problem, generally known as the label-switching problem. A model with a given sequence of component parameters is observationally equivalent to any permutations of this sequence of parameters. Component labels, therefore, require identifying restrictions for inference to occur. One solution to this problem is to put informative priors on the model parameters (e.g., $\bar{\theta}_1 > \bar{\theta}_2 > \cdots > \bar{\theta}_K$), which works well when the data are in agreement with the restriction. However, if the data are not in agreement (e.g., the components primarily differ in V, not $\bar{\theta}$), then the prior can lead to a chain that is slow to converge (Frühwirth-Schnatter et al. 2003). It should be noted, however, that the presence of label-switching does not affect inference about parameters of a particular unit, θ_i. If the normal component mixing distribution is seen as a flexible device for approximating some unknown heterogeneity distribution, then inference about the distribution of heterogeneity can be made directly with the set of unit parameters, $\{\theta_i\}$, without attempting to identify or estimate the component parameters.

In many situations, we have prior information on the signs of various coefficients in the base model. For example, price parameters are negative and advertising effects are positive. In a Bayesian approach, this sort of prior information can be included by modifying the first-stage prior. We replace the normal distribution with a distribution with restricted support, corresponding to the appropriate sign restrictions. For example, we can use a log-normal distribution for a parameter which is restricted via sign by the reparameterization, $\theta' = \ln(\theta)$. However, note that this change in the form of the prior can destroy some of the conjugate relationships which are exploited in the Gibbs-sampler. However, if metropolis-style methods are used to generate draws in the Markov chain, it is a simple matter to directly reparameterize the likelihood function, by substituting $\exp(\theta')$ for θ, rather than rely on the heterogeneity distribution to impose the range restriction. What is more important is to ask whether the log-normal prior is appropriate. The left tail of the log-normal distribution declines to zero, insuring a mode for the log-normal distribution at a strictly positive value. For situations in which we want to admit zero as a possible value for the parameter, this prior may not be appropriate. Boatwright et al. (1999) explore the use of truncated normal priors as an alternative to the log-normal reparameterization approach. Truncated normal priors are much more flexible, allowing for mass to be piled up at zero.

Bayesian models can also accommodate structural heterogeneity, or changes in the likelihood specification for a unit of analysis. The likelihood is specified as a mixture of likelihoods:

$$p(y_{it} \mid \{\theta_{ik}\}) = r_1 p_1(y_{it} \mid \theta_{i1}) + \cdots + r_K p_K(y_{it} \mid \theta_{iK}),$$

and estimation proceeds by appending indicator variables for the mixture component to the state space. Conditional on the indicator variables, the datum, y_{it}, is assigned to one of K likelihoods. The indicator variables, conditional on all other parameters, have a multinomial distribution with probabilities proportional to the number of observations assigned to the component, and the probability that the datum arise from likelihood. Models of structural heterogeneity have been used to investigate intraindividual change in the decision process due to environmental changes (Yang and Allenby 2000) and fatigue (Otter et al. 2003).

Finally, Bayesian methods have recently been used to relax the commonly made assumption that the unit parameters, θ_i, are i.i.d. draws from the distribution of heterogeneity. Ter Hofstede et al. (2002) employ a conditional Gaussian field specification to study spatial patterns in response coefficients:

$$p(\theta_i \mid \tau) = p(\theta_i \mid \{\theta_j \colon j \in S_i\}, V_\theta),$$

where S_i denotes units that are spatially adjacent to unit i. Since the MCMC estimation algorithm employs full conditional distributions of the model parameters, the draw of θ_i involves using a local average for the mean of the mixing distribution. Yang and Allenby (2002b) employ a simultaneous specification of the unit parameters to reflect the possible presence of interdependent effects, due to the presence of social and information networks.

$$\theta = \rho W\theta + u$$
$$u \sim N(0, \sigma^2 I),$$

where W is a matrix that specifies the network, ρ, is a coefficient that measures the influence of the network, and u is an innovation.

5.5. Diagnostic Checks of the First-Stage Prior

In the hierarchical model, the prior is specified in a two stage process:

$$\theta \sim N(\bar{\theta}, V_\theta)$$
$$p(\bar{\theta} V_\theta).$$

In the classical literature, the normal distibution of θ would be called the random effects model and would be considered part of the likelihood, rather than part of the prior. Typically, very diffuse priors are used for the second stage. Thus, it is the first-stage prior which is important, and will always remain important, as long as there are only a few observations available per household. Since the parameters of the first-stage prior are inferred from the data, the main focus of concern should be on the form of this distribution.

In the econometric literature, the use of parametric distributions of heterogeneity (e.g., normal distributions) are often criticized on the grounds that their misspecification leads to inconsistent estimates of the common model parameters (cf. Heckman and Singer 1982). For example, if the true distribution of household parameters were skewed or bimodal, our inferences based on a symmetric, unimodal normal prior could be misleading. One simple approach would be to plot the distribution of the posterior household means and compare this to the implied normal distribution evaluated at the Bayes estimates of the hyperparameters, $N(E[\bar{\theta} \mid \text{data}], E[V_\theta])$. The posterior means are not constrained to follow the normal distribution because the normal distribution is only part of the prior and the posterior is influenced by the unit-level data. This simple approach is in the right spirit but could be misleading due to the fact that we do not properly account for uncertainty in the unit-level parameter estimates.

Allenby and Rossi (1999) provide a diagnostic check of the assumption of normality in the first stage of the prior distribution that properly accounts for parameter uncertainty. To handle uncertainty in our knowledge of the common parameters of the normal distribution, we compute the predictive distribution of $\theta_{i'}$ for unit i', selected at random from the population of households with the random effects distribution. Using our data and model, we can define the predictive distribution of $\theta_{i'}$ as follows:

$$\theta_{i'} \mid \text{data} \; = \iint \phi(\theta \mid \bar{\theta}, V_\theta) p(\bar{\theta}, V_\theta \mid \text{data})\, d\bar{\theta}\, dV_\theta.$$

Here $\phi(\theta_{i'} \mid \bar{\theta}, V_\theta)$ is the normal prior distribution. We can use our MCMC draws of $\bar{\theta}$, V_θ, coupled with draws from the normal prior, to construct an estimate of this distribution. The diagnostic check is constructed by comparing the distribution of the unit-level posterior means to the predictive distribution based on the model given above.

5.6. Findings and Influence on Marketing Practice

The last ten years of work on heterogeneity in marketing has yielded several important findings.

Researchers have explored a rather large set of first-stage models with a normal distribution of heterogeneity across units. In particular, investigators have considered a first-stage normal linear regression (Blattberg and George 1991), a first-stage logit model (Allenby and Lenk 1994, 1995), a first-stage probit (McCulloch and Rossi 1994), a first-stage Poisson (Neelamegham and Chintagunta 1999), and a first-stage generalized gamma distribution model (Allenby et al. 1999, Jen et al. 2003). The major conclusion is that there is a substantial degree of heterogeneity across units in various marketing data sets. This finding of a large degree of heterogeneity holds out substantial promise for the study of preferences, both in terms of substantive and practical significance (Ansari et al. 2000). There may be substantial heterogeneity bias in models that do not properly account for heterogeneity (Chang et al. 1999), and there is large value in customizing marketing decisions to the unit level (see Rossi et al. 1996).

Yang et al. (2002a) investigate the source of brand preference, and find evidence that variation in the consumption environment, and resulting motivations, leads to changes in a unit's preference for a product offering (see also, Arora and Allenby 1999). Motivating conditions are an interesting domain for research, as they preexist the marketplace, offering a measure of demand that is independent of marketplace offerings. Other research has documented evidence that the decision process employed by a unit is not necessarily constant throughout a unit's purchase (Yang and Allenby 2000) and response (Otter et al. 2003) history. This evidence indicates that the appropriate unit of analysis for marketing is at the level that is less aggregate than a person or respondent, although there is evidence that household sensitivity to marketing variables (Ainslie and Rossi 1998) and state dependence (Seetharaman et al. 1999) is constant across categories.

The normal continuous model of heterogeneity appears to do reasonably well in characterizing this heterogeneity, but there has not yet been sufficient experimentation with alternative models, such as the mixture of normals, to draw any definitive conclusions (see Allenby et al. 1998). With the relatively short panels typically found in marketing applications, it may be difficult to identify much more detailed structure beyond that afforded by the normal model. In addition, relatively short panels may produce a confounding of the finding of heterogeneity with various model misspecifications in the first stage. If only one observation is available for each unit, then the probability model for the unit level is the mixture of the first-stage model with the second-stage prior:

$$p(y \mid \tau) = \int p(y \mid \theta) p(\theta \mid \tau)\, d\theta.$$

This mixing can provide a more flexible probability model. In the one observation situation, we can never determine whether it is "heterogeneity," or lack of flexibility that causes the Bayesian hierarchical model to fit the data well. Obviously, with more than one observation per unit, this changes, and it is possible to separately diagnose first-stage model problems and deficiencies in the assumed heterogeneity distribution. However, with short panels there is unlikely to be a clean separation between these problems, and it may be the case that some of the heterogeneity detected in marketing data is really due to lack of flexibility in the base model.

There have been some comparisons of the normal continuous model with the discrete approximation approach of a finite-mixture model. It is our view that it is conceptually inappropriate to view any population of units as being comprised of only a small number of homogeneous groups and, therefore, the appropriate interpretation of the finite mixture approach is an approximation method. Allenby and Rossi (1999) and Lenk et al. (1996) show some of the shortcomings of the finite-mixture model, and provide some evidence that the finite-mixture model does not recover reasonable unit-level parameter estimates. In contrast, Andrews et al. (2002) use simulated data to suggest that unit-level recovery is comparable between the normal- and finite-mixture approaches.

At the same time that the Bayesian work in the academic literature has shown the ability to produce unit-level estimates, there has been increased interest on the part of practitioners in unit-level analysis. Conjoint researchers have always had an interest in respondent-level part-worths and had various ad hoc schemes for producing these estimates. Recently, the Bayesian hierarchical approach to the logit model has been implemented in the popular Sawtooth conjoint software. Experience with this software and simulation studies have lead Rich Johnson, Sawtooth software's founder, to conclude that Bayesian methods are superior to others considered in the conjoint literature (Sawtooth Software 2001).

Retailers are amassing volumes of store-level scanner data. Not normally available to academic researchers, this store-level data is potentially useful for informing the basic retail decisions such as pricing and merchandizing. Attempts to develop reliable models for pricing and promotion have been frustrated by the inability to produce reliable promotion and price response parameters. Thus, the promise of store-level pricing has gone unrealized. Recently, a number of firms, including the leader DemandTec, have appeared in this space, offering data-based pricing and promotion services to retail customers. At the heart of DemandTec's approach is a Bayesian shrinkage model applied to store-sku-week data, obtained directly from the retail client. The Bayesian shrinkage methods allow DemandTec to produce reasonable and relatively stable store-level parameter estimates. DemandTec builds on the approach of Montgomery (1997).

6. Decision Theory

The vast majority of the recent Bayesian literature in marketing emphasizes the value of the Bayesian approach to inference, particularly in situations with limited information. Bayesian inference is only a special case of the more general Bayesian decision theoretic approach. Bayesian decision theory has two critical and separate components: (1) a loss function, and (2) the posterior distribution. The loss function associates a loss with a state of nature and a action, $l(a, \theta)$, where a is the action and θ is the state of nature (parameter). The optimal decision maker chooses the action so as to minimize expected loss, where the expectation is taken with respect to the posterior distribution.

$$\min_a \bar{l}(a) = \int l(a, \theta) p(\theta \mid \text{data})\, d\theta.$$

Parameter inference is a simple case of the general decision theory set-up, in which the loss is often taken to be quadratic. In this case, the optimal "action" is an estimator taken to be the posterior mean of the parameters.

6.1. Model Selection

In many scientific settings, the action is a choice between competing models. In the Bayesian approach, it is possible to define a set of models $M_1, \ldots, M_k$, and calculate a measure of the posterior probability of a model. If the loss function is zero when the correct model is chosen and equal for all cases in which the incorrect model is chosen, then the optimal Bayesian decision maker chooses the model with the highest posterior probability. In a parametric setting, the posterior probability of a model can be calculated as follows:

$$p(M_k \mid D) = p(D \mid M_k) p(M_k)$$
$$p(D \mid M_k) = \int p(D \mid \theta, M_k) p_k(\theta)\, d\theta,$$

where D denotes the "data." In the Bayesian approach, the posterior probability only requires specification of the class of models and the priors. There is no distinction between nested and nonnested models as in the hypothesis-testing literature in the classical literature. However, we do require specification of the class of models under consideration; there is no omnibus measure of the plausibility of a given model or group of models versus some unspecified, and possibly unknown, set of alternative models.

In situations where two models are being compared, it is common to compute the ratio of posterior model probabilities. This ratio can be expressed as the ratio of average likelihoods times the prior odds ratio. The ratio of average likelihood is sometimes called the Bayes factor for a model.

$$\frac{p(M_1 \mid D)}{p(M_2 \mid D)} = \frac{\int l_1(\theta_1) p_1(\theta_1)\, d\theta_1}{\int l_2(\theta_2) p_2(\theta_2)\, d\theta_2} \times \frac{p(M_1)}{p(M_2)}.$$

The Bayes factor can be quite sensitive to the prior specification and, in particular, to the prior diffusion. As the prior becomes more and more spread out, relative to the fixed likelihood, the average value of the likelihood declines. Thus, if the prior for Model 1 is a great deal more spread out than the prior for Model 2, this may result in Bayes factors which favor Model 2 (this is certainly true in a limiting sense). In particular, diffuse and improper priors can result in undefined Bayes factors. We recommend that close attention be placed on the prior assessment and that prior sensitivity analysis be performed whenever computing posterior model probabilities.

A wide variety of methods have been proposed to approximate the posterior model probability. The most widely used method is due to Schwarz (1978), who computed an asymptotic approximation that depends only on the dimension of the model. This is the idea behind the well-known Schwarz or Bayesian Information Criterion (BIC) for model choice. Except for very special forms of priors, the Schwarz method is extremely inaccurate and should not be relied on for computation of the posterior model probability. Various numerical methods that rely on either the Laplace approximation or importance sampling methods of numerical integration are the preferred method of approximation. In particular, Newton and Raftery (1994) offer a convenient method for approximating a Bayes factor using MCMC simulation draws to estimate the average likelihood as the harmonic mean of the likelihoods of a sample from the posterior distribution. This estimator is consistent but may be unstable due to draws of the parameters that are associated with small likelihood values.

6.2. Marketing Decisions and Bayesian Decision Theory

Bayesian decision theory is ideally suited for application to many marketing problems in which a decision must be made, given substantial parameter or modeling uncertainty. In these situations, the uncertainty must factor into the decision itself. The marketing decision maker takes an action by setting the value of various variables designed to quantify the marketing environment facing the consumer (such as price or advertising levels). These decisions should be affected by the level of uncertainty facing the marketer. To make this concrete, begin with a probability model that specifies how the outcome variable (y) is driven by the explanatory variables (x) and parameters θ.

$$p(y \mid x, \theta).$$

The decision maker has control over a subset of the x vector, $x' = [x'_d, x'_{\text{cov}}]$. x_d represents the variables under the decision maker's control and x_{cov} are the covariates. The decision maker chooses x_d so as to maximize the expected value of profits where the expectation is taken over the distribution of the outcome variable. In a fully Bayesian decision theoretic treatment, this expectation is taken with respect to the posterior distribution of θ, as well as the predictive conditional distribution $p(y \mid x_d, x_{\text{cov}})$.

$$\begin{aligned}\pi^*(x_d \mid x_{\text{cov}}) &= E_\theta[E_{y\mid\theta}[\pi(y \mid x_d)]]\\ &= E_\theta\left[\int \pi(y \mid x_d) p(y \mid x_d, x_{\text{cov}}, \theta)\, dy\right]\\ &= E_\theta[\bar{\pi}(x_d \mid x_{\text{cov}}, \theta)].\end{aligned}$$

The decision maker chooses x_d to maximize profits π^*. In general, the decision maker can be viewed as minimizing expected loss, which is frequently taken as profits but need not be in all cases (see, for example, Steenburgh et al. 2002)

6.3. Plug-In vs. Full Bayes Approaches

The use of the posterior distribution of the model parameters to compute expected profits is an important aspect of the Bayesian approach. In an approximate, or conditional, Bayes approach, the integration of the profit function with respect to the posterior distribution of θ is replaced by an evaluation of the function at the posterior mean or mode of the parameters. This approximate approach is often called the "plug-in" approach, or according to Morris (1983), Bayes Empirical Bayes.

$$\pi^*(x_d) = E_{\theta|y}[\bar{\pi}(x_d \mid \theta)] \neq \bar{\pi}(x_d \mid \hat{\theta} = E_{\theta|y}[\theta]).$$

When the uncertainty in θ is large and the profit function is nonlinear, errors from the use of the plug-in method can be large. In general, failure to account for parameter uncertainty will overstate the potential profit opportunity and lead to "overconfidence" that results in an overstatement of the value of information (see also Allenby 1990b, Kalyanam 1996, Montgomery and Bradlow 1999).

6.4. Use of Alternative Information Sets

One of the most appealing aspects of the Bayesian approach is the ability to incorporate a variety of different sources of information. All adaptive shrinkage methods utilize the similarity between cross-sectional units to improve inference at the unit level. A high level of similarity among units leads to a high level of information shared. Because the level of similarity is determined by the data via the first-stage prior, the shrinkage aspects of the Bayesian approach adapt to the data. For example, Neelameghan and Chintagunta (1999) show that similarities between countries can be used to predict the sales patterns following the introduction of new products.

The value of a given information set can be assessed using a profit metric and the posteriors of θ, corresponding to the two information sets. For example, consider two information sets A and B, along with corresponding posteriors, $p_A(\theta)$, $p_B(\theta)$. We solve the decision problem using these two posterior distributions.

$$\Pi_l = \max_{x_d} \pi_l^*(x_d \mid x_{\text{cov}}) = \max_{x_d} \int \bar{\pi}(x_d \mid x_{\text{cov}}, \theta) p_l(\theta)\, d\theta \qquad l = A, B.$$

Rossi et al. (1996) use this approach to value various information sets available on individual households. A targeting couponing problem that anticipated the now popular Catalina Marketing Inc. products was used to value a sequence of expanding individual level information sets. We now turn to the problem of valuing disaggregate information.

6.5. Valuation of Disaggregate Information

Once a fully decision-theoretic approach has been specified, we can use the profit metric to value the information in disaggregate data. We compare profits that can be obtained via our disaggregate inferences about $\{\theta_i\}$ with profits that could be obtained using only aggregate information. The profit opportunities afforded by disaggregate data will depend on both the amount of heterogeneity across the units in the panel data, as well as the level of information at the disaggregate level.

To make these notions explicit, we will lay out the disaggregate and aggregate decision problems. As emphasized in §3, Bayesian methods are ideally suited for inference about the individual or disaggregate parameters, as well as the common parameters. Recall the profit function for the disaggregate decision problem.

$$\pi_i^*(x_{d,i} \mid x_{\text{cov},i}) = \int \bar{\pi}(x_{d,i} \mid x_{\text{cov},i}, \theta_i) p(\theta_i \mid \text{data})\, d\theta_i.$$

Here, we take the expectation with respect to the posterior distribution of the parameters for unit "i." Total profits from the disaggregate data are simply the sum of the maximized values of the profit function above.

$$\Pi_{\text{disagg}} = \sum \pi_i^*(\tilde{x}_{d,i} \mid x_{\text{cov},i}) \quad \text{where } \tilde{x}_{d,i} \text{ is the optimal choice of } x_{d,i}.$$

Aggregate profits can be computed by maximizing the expectation of the sum of the disaggregate profit functions with respect to the predictive distribution of θ_i

$$\begin{aligned}\pi_{\text{agg}}(x_d) &= E_\theta\Big[\sum \bar{\pi}(x_d \mid x_{\text{cov},i}, \theta)\Big] \\ &= \int \sum \bar{\pi}(x_d \mid x_{\text{cov},i}, \theta) \bar{p}(\theta)\, d\theta \\ \Pi_{\text{agg}} &= \pi_{\text{agg}}(\tilde{x}_d).\end{aligned}$$

The appropriate predictive distribution of θ, $\bar{p}(\theta)$, is formed from the marginal of the first-stage prior with respect to the posterior distribution of the model parameters.

$$\bar{p}(\theta) = \int p(\theta \mid \tau) p(\tau \mid \text{data})\, d\tau.$$

Comparison of Π_{agg} with Π_{disagg} provides a metric for the achievable value of the disaggregate information.

7. Open Issues and Directions for Future Research

Researchers have long noted the conceptual appeal of the Bayesian framework for inference and decision making. However, the potential of the Bayesian approach was not realized due to computational constraints. Without modern simulation-based methods, researchers were restricted to a short list of likelihoods and associated conjugate priors. The developments of the last 15 years have freed us from computation constraints, allowing for the analysis of virtually any model. We now can consider models once thought to be impossible to compute, and we can use priors of virtually any form. The only constraint now, is the ability of the data to identify model parameters, rather than the ability of the analyst to conduct inference for this model. However, the recent developments have an even more profound impact than simply freeing us from computational constraints. The nature of the MCMC methods emphasize a modularity in the construction of models, typically achieved through a combination of conditional distributions. These conditional distributions specify the nature of the relationships between observed variables and allow for the construction of more complicated relationships. Thus, the researcher can create a more complex model simply by adding layers to the hierarchy.

Consider, as a simple example, the relationship between sales and price. Much attention has been devoted to fitting the conditional distribution of sales (y) given price (x). However, the actual decision process is certainly not well represented by one conditional distribution. Many endorse the concept of a latent consideration set (Chiang et al. 1999) in which a product must first be included in the consideration set before a consumer evaluates the impact of price. If w represents the consideration set, then the model has been enlarged to the two layers $y \mid w, x$, and $w \mid z$, where the consideration set is influenced by another variable z (e.g., advertising). In the end, the hierarchical model specifies a special form for the conditional distribution of $y \mid x, z$ that allows exploration of the intermediary conditional relationships. Moreover, the specification of hierarchical conditional models is consistent with process models of consumer behavior (e.g., McFadden 2001).

Consideration sets are only one example of a latent process that intervenes between the measurements of the marketing mix variables and the sales outcome variable. Other important examples include price search and consumption. In typical demand data, we do not observe the consumption of goods but merely their purchases. In much demand modeling in marketing, this distinction is glossed over, and the demand model is based on a utility function defined directly on the purchase quantities. Models that explicitly recognize that purchases are made in anticipation of future consumption have recently received attention. For example, Dube (2003) explains simultaneous purchases of different varieties via anticipation of changes in tastes over future consumption occasions. Yang et al. (2002) consider a model in which the utility derived from goods is dependent on the context of consumption. Erdem and Keane (2003) consider dynamic models of consumer demand in which households stockpile goods for future consumption. All of these models are amenable to Bayesian analysis via data augmentation in which latent variables, such as consumption, are introduced into the inference procedures.

Price search models are another example of a latent process of great importance in marketing. Consumers are not always fully informed about the prices of choice alternatives and must engage in price search. We do not observe this price search process directly but only the outcomes. In a classical approach, such as Mehta and Srinivasan (2003), the likelihood for the search model must be evaluated by integrating over all possible search paths. In a data augmentation approach, this integration can be achieved by introduction of latent variables that represent search possibilities. In an MCMC method for navigating the posterior distribution of search parameters and latent variables, we do not enumerate all possible search paths but, instead, navigate among paths of high posterior probability. We believe that MCMC approaches, together with data augmentation, hold great promise for analyzing models with very large latent state spaces such as price search models and discrete dynamic programming models, in general.

Many models of consumer behavior include threshold-like effects. For example, some models of consideration set formation have screening rules in which a threshold level of an attribute is defined. The threshold levels are unobservable parameters, and the likelihood over these parameters has discontinuities. This rules out the use of standard derivative-based maximization methods. MCMC methods simply require draws from various conditional posterior distributions in order to navigate the parameter space. Drawing from a distribution with a density that is not continuous poses no special difficulties. Gilbride and Allenby (2003) illustrate how this can be implemented for choice models with conjunction and disjunctive screening rules. These developments open many possibilities for analysis of models with threshold components.

Thus, hierarchical modeling methods achieve not only a great flexibility as emphasized in the Bayesian statistics literature, but also they are well-suited to the elaboration of various latent process views of consumer behavior and decision making. We expect research in marketing to focus on a better understanding of the process by which the consumer makes buying decisions, in hopes of creating more realistic, yet still parsimonious, models of behavior.

A major challenge facing marketing practitioners is the merging of information acquired across a variety of different datasets. For example, a firm may have access to consumer purchase information, survey information on a subsample of consumers, and syndicated aggregate sales information. Marketplace and survey data cannot be combined without some view to the processes by which consumers make buying decisions and respond to survey instruments. Bayesian methods will facilitate the integration of these data sources through the specification of a common set of behavioral parameters and the processes by which these are translated into either survey responses or purchase decisions.

The observational data used in much of quantitative marketing is derived from an environment in which the outcome and input variables are jointly determined. Marketing mix variables are set by managers with a view toward optimizing some objective function that includes the dependent variable. For example, prices may be set with some knowledge of either price sensitivity or price demand shocks. Direct marketing response data is obtained from samples of consumers who were selected in a nonrandom fashion, with a view toward maximizing response rates or profitability. Sales forces are allocated using some sort of heuristic that attempts to create an optimal allocation in which the marginal benefit of further effort is equated to marginal cost. This means that we cannot model just the conditional distribution of the outcome variable, given the marketing mix variables, but that we must consider the joint distribution of all variables.

The joint determination of both outcome and input variables poses considerable challenges for statistical inference and modeling. Manchanda et al. (2003) consider sales force problems in which the level of sales force effort at a given account is a function of sales response parameters. Price endogeneity is another example of a challenging problem that involves deriving the joint distribution of price, sales, and possible exogeneous variables. Computational difficulties have limited the use of likelihood-based methods and, instead, instrumental variables procedures have been commonly employed. We believe there is substantial room for improvement in this area by the use of likelihood-based Bayesian approaches. As an example, consider a model of demand and supply in which there are cost shocks and a common demand shock that is used by retailers in setting prices. This model has a likelihood that is the joint distribution of price and quantity sold. This joint distribution is derived from the distribution of costs shocks and demand shocks. While the mapping from shocks to observables is an implicit nonlinear system of equations, there is no conceptual difficulty with implementing a metropolis algorithm for this system. The modularity of the metropolis style MCMC method means that elaborating the model by adding, for example, consumer heterogeneity, is straightforward (see Yang et al. 2003).

8. Conclusion

We have emphasized the value of Bayesian methods in situations with limited information. While the total amount of data available has exploded, the amount of information about any one consumer is likely to remain limited. The customization of marketing actions to finer and finer levels of aggregation requires the ability to make inferences in conditions of limited information and to characterize the level of uncertainty in these inferences. Thus, we expect Bayesian methods will play a critical role in realizing the potential of micromarketing and any analysis conducted at a microlevel.

Finally, there are a number of important problems in marketing that are essentially pure prediction problems. Given a set of information on a consumer, the prediction problem is to predict the response to a given configuration of the marketing environment. Information available about the consumer can be summarized with a huge set of potential variables. The marketing environment itself can also be summarized in many possible ways. One important applied problem is to sift through a large number of possible variables and functional forms to find the best possible prediction rule. In the Bayesian statistics literature, there has been substantial progress in the "variable selection" problem, and we believe these methods have great promise for application to marketing problems.

Structural or process-oriented approaches to modeling achieve the prediction goal via a specification of the decision process. This guides in the selection of variables and in the structure of relationships between variables. However, structural theories are typically silent on the exact parametric form of functional relationships or distributions. Again, there is an opportunity for application of Bayesian nonparametric methods to the structural approach as well (Kalyanam and Shively 1998, Shively et al. 2000).

In summary, Bayesian statistical methods offer an appealing set of tools to researchers in marketing. The Bayesian approach offers an integrated view of inference and decision making that is applicable to both theoretical and applied analysis. Moreover, the hierarchical modeling structure that is exploited in MCMC estimation methods is congruent with theories of behavior and offers a means of integrating information across multiple data sources. Finally, the computational advantages of Bayesian methods allow for study of high-dimensional data and complex relationships that are common in marketing. We encourage our colleagues and students to experiment with and apply Bayesian methods.

Acknowledgments

Author Peter Rossi thanks the James M. Kilts Center for Marketing support of this research. The authors thank Rob McCulloch and Eric Bradlow for useful comments. The authors were inspired by the writings and teachings of Arnold Zellner and Dennis Lindley throughout their careers.

Appendix: Annotated Citations of Bayesian Applications in Marketing

This annotated bibliography represents the results of a search for applications of Bayesian statistics in marketing. Only published or forthcoming articles that feature marketing applications are included.

Ainslie, Andrew, and Peter Rossi. 1998. Similarities in choice behavior across product categories. *Marketing Sci.* **17** 91–106.

A multi-category choice model is proposed where household response coefficients are assumed dependent across category. The estimated distribution of heterogeneity reveals that price, display, and feature sensitivity are not uniquely determined for each category but may be related to household-specific factors.

Allenby, Greg M., Thomas Shively, Sha Yang, Mark J. Garratt. 2003. A choice model for packaged goods: Dealing with discrete quantities and quantity discounts. *Marketing Sci.* Forthcoming.

A method for dealing with the pricing of a product with different package sizes is developed from utility-maximizing principles. The model allows for the estimation of demand when there exist a multitude of size-brand combinations.

Allenby, Greg M., Robert P. Leone, Lichung Jen. 1999. A dynamic model of purchase timing with application to direct marketing. *J. Amer. Statist. Assoc.* **94** 365–374.

Customer interpurchase times modeled with a heterogeneous generalized gamma distribution, where the distribution of heterogeneity is a finite mixture of inverse generalized gamma components. The model allows for structural heterogeneity where customers can become inactive.

Allenby, Greg M., Neeraj Arora, James L. Ginter. 1998. On the heterogeneity of demand. *J. Marketing Res.* **35** 384–389.

A normal component mixture model is compared to a finite mixture model using conjoint data and scanner panel data. The predictive results provide evidence that the distribution of heterogeneity is continuous, not discrete.

Allenby, Greg M., Lichung Jen, Robert P. Leone. 1996. Economic trends and being trendy: The influence of consumer confidence on retail fashion sales. *J. Bus. Econom. Statist.* **14** 103–111.

A regression model with autoregressive errors is used to estimate the influence of consumer confidence on retail sales. Data are pooled across divisions of a fashion retailer to estimate a model where influence has a differential impact on preseason versus in-season sales.

Allenby, Greg M., Peter J. Lenk. 1995. Reassessing brand loyalty, price sensitivity, and merchandising effects on consumer brand choice. *J. Bus. Econom. Statist.* **13** 281–289.

The logistic normal regression model of Allenby and Lenk (1994) is used to explore the order of the brand-choice process and to estimate the magnitude of price, display, and feature advertising effects across four scanner panel datasets. The evidence indicates that brand-choice is not zero order, and merchandising effects are much larger than previously thought.

Allenby, Greg M., James L. Ginter. 1995. Using extremes to design products and segment markets. *J. Marketing Res.* **32** 392–403.

A heterogeneous random-effects binary choice model is used to estimate conjoint part-worths using data from a telephone survey. The individual-level coefficients available in hierarchical Bayes models are used to explore extremes of the heterogeneity distribution, where respondents are most and least likely to respond to product offers.

Allenby, Greg M., Neeraj Arora, James L. Ginter. 1995. Incorporating prior knowledge into the analysis of conjoint studies. *J. Marketing Res.* **32** 152–162.

Ordinal prior information is incorporated into a conjoint analysis using a rejection sampling algorithm. The resulting part-worth estimates have sensible algebraic signs that are needed for deriving optimal product configurations.

Allenby, Greg M., Peter J. Lenk. 1994. Modeling household purchase behavior with logistic normal regression. *J. Amer. Statist. Assoc.* **89** 1218–1231.

A discrete choice model with autocorrelated errors and consumer heterogeneity is developed and applied to scanner panel dataset of ketchup purchases. The results indicate substantial unobserved heterogeneity and autocorrelation in purchase behavior.

Allenby, Greg M. 1990a. Hypothesis testing with scanner data: The advantage of Bayesian methods. *J. Marketing Res.* **27** 379–389.

Bayesian testing for linear restrictions in a multivariate regression model is developed and compared to classical methods.

Allenby, Greg M. 1990b. Cross-validation, the Bayes theorem, and small-sample bias. *J. Bus. Econom. Statist.* **8** 171–178.

Cross-validation methods that employ plug-in point approximations to the average likelihood are compared to formal Bayesian methods. The plug-in approximation is shown to overstate the amount of statistical evidence.

Andrews, Rick, Asim Ansari, Imran Currim. 2002. Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *J. Marketing Res.* 87–98.

A simulation study is used to investigate the performance of continuous and discrete distributions of heterogeneity in a regression model. The results indicate that Bayesian methods are robust to the true underlying distribution of heterogeneity, and finite mixture models of heterogeneity perform well in recovering true parameter estimates.

Ansari, Asim., Skander Essegaier, Rajeev Kohli. 2000. Internet recommendation systems. *J. Marketing Res.* **37** 363–375.

Random-effect specifications for respondents and stimuli are proposed within the same linear model specification. The model is used to pool information from multiple data sources.

Ansari, Asim, Kamel Jedidi, Sharan Jagpal. 2000. A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Sci.* **19** 328–347.

Covariance matrix heterogeneity is introduced into a structural equation model, in contrast to standard models in marketing, where heterogeneity is introduced into the mean structure of a model. The biasing effects of not accounting for covariance heterogeneity are documented.

Arora, Neeraj, Greg M. Allenby. 1999. Measuring the influence of individual preference structures in group decision making. *J. Marketing Res.* **36** 476–487.

Group preferences differ from the preferences of individuals in the group. The influence of the group on the distribution of heterogeneity is examined using conjoint data on durable good purchases by a husband's, a wife's, and their joint evaluation.

Arora, Neeraj, Greg M. Allenby, James L. Ginter. 1998. A hierarchical Bayes model of primary and secondary demand. *Marketing Sci.* **17** 29–44.

An economic discrete/continuous demand specification is used to model volumetric conjoint data. The likelihood function is structural, reflecting constrained utility maximization.

Blattberg, Robert C., Edward I. George. 1991. Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *J. Amer. Statist. Assoc.* **86** 304–315.

Weekly sales data across multiple retailers in a chain are modeled using a linear model with heterogeneity. Price and promotional elasticity estimates are shown to have improved predictive performance.

Boatwright, Peter, Robert McCulloch, Peter E. Rossi. 1999. Account-level modeling for trade promotion: An application of a constrained parameter hierarchical model. *J. Amer. Statist. Assoc.* **94** 1063–1073.

A common problem in the analysis of sales data is that price coefficients are often estimated with algebraic signs that are incompatible with economic theory. Ordinal constraints are introduced through the prior to address this problem, leading to a truncated distribution of heterogeneity.

Bradlow, Eric T., David Schmittlein. 1999. The little engines that could: Modeling the performance of World Wide Web search engines. *Marketing Sci.* **19** 43–62.

A proximity model is developed for analysis of the performance of Internet search engines. The likelihood function reflects the distance between the engine and specific URLs, with the mean location of the URLs parameterized with a linear model.

Bradlow, Eric T., S. Fader. 2001. A Bayesian lifetime model for the "Hot 100" *Billboard* songs. *J. Amer. Statist. Assoc.* **96** 368–381.

A time series model for ranked data is developed using a latent variable model. The deterministic portion of the latent variable follows a temporal pattern described by a generalized gamma distribution, and the stochastic portion is extreme value.

Bradlow, Eric T., Vithala R. Rao. 2000. A hierarchical Bayes model for assortment choice. *J. Marketing Res.* **37** 259–268.

A statistical measure of attribute assortment is incorporated into a random-utility model to measure consumer preference for assortment beyond the effects from the attribute levels themselves. The model is applied to choices between bundled offerings.

Chiang, Jeongwen, Siddartha Chib, Chakravarthi Narasimhan. 1999. Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. *J. Econometrics* **89** 223–248.

Consideration sets are enumerated and modeled with a Dirichlet prior in a model of choice. A latent state variable is introduced to indicate the consideration set, resulting in a model of structural heterogeneity.

Chang, Kwangpil, S. Siddarth, Charles B. Weinberg. 1999. The impact of heterogeneity in purchase timing and price responsiveness on estimates of sticker shock effects. *Marketing Sci.* **18** 178–192.

A random utility model with reference prices is examined, with and without allowance for household heterogeneity. When heterogeneity is present in the model, the reference price coefficient is estimated to be close to zero.

DeSarbo, Wayne, Youngchan Kim, Duncan Fong. 1999. A Bayesian multidimensional scaling procedure for the spatial analysis of revealed choice data. *J. Econometrics* **89** 79–108.

The deterministic portion of a latent variable model is specified as a scalar product of consumer and brand coordinates to yield a spatial representation of revealed choice data. The model provides a graphical representation of the market structure of product offerings.

Edwards, Yancy, Greg M. Allenby. 2003. Multivariate analysis of multiple response data. *J. Marketing Res.* Forthcoming.

Pick any of *J* data is modeled with a multivariate probit model, allowing standard multivariate techniques to be applied to the parameter of the latent normal distribution. Identifying restrictions for the model are imposed by post-processing the draws of the Markov chain.

Huber, Joel, Kenneth Train. 2001. On the similiarity of classical and Bayesian estimates of individual mean partworths. *Marketing Lett.* **12** 259–269.

Classical and Bayesian estimation methods are found to yield similar individual-level estimates. The classical methods condition on estimated hyperparameters, while Bayesian methods account for their uncertainty.

Jen, Lichung, Chien-Heng Chou, Greg M. Allenby. 2003. A Bayesian approach to modeling purchase frequency. *Marketing Lett.* **14** 5–20.

A model of purchase frequency that combines a Poisson likelihood with gamma mixing distribution is proposed, where the mixing distribution is a function of covariates. The covariates are shown to be useful for customers with short purchase histories or have infrequent interaction with the firm.

Kalyanam, Kirthi, Thomas S. Shively. 1998. Estimating irregular pricing effects: A stochastic spline regression approach. *J. Marketing Res.* **35** 16–29.

Stochastic splines are used to model the relationship between price and sales, resulting in a more flexible specification of the likelihood function.

Kalyanam, Kirthi. 1996. Pricing decision under demand uncertainty: A Bayesian mixture model approach. *Marketing Sci.* **15** 207–221.

Model uncertainty is captured in model predictions by taking a weighted average where the weights correspond to the posterior probability of the model. Pricing decisions are shown to be more robust.

Kamakura, Wagner A., Michel Wedel. 1997. Statistical data fusion for cross-tabulation. *J. Marketing Res.* **34** 485–498.

Imputation methods are proposed for analyzing cross-tabulated data with empty cells. Imputation is conducted in an iterative manner to explore the distribution of missing responses.

Kim, Jaehwan, Greg M. Allenby, Peter E. Rossi. 2002. Modeling consumer demand for variety. *Marketing Sci.* **21** 223–228.

A choice model with interior and corner solutions is derived from a utility function with decreasing marginal utility. Kuhn-Tucker conditions are used to relate the observed data, with utility maximization in the likelihood specification.

Lee, Jonathan, Peter Boatwright, Wagner Kamakura. 2003. A Bayesian model for prelaunch sales forecasting of recorded music. *Management Sci.* **49** 179–196.

The authors study the forecasting of sales for new music albums prior to their introduction. A hierarchical logistic shaped diffusion model is used to combine a variety of sources of information on attributes of the album, effects of marketing variables, and dynamics of adoption.

Leichty, John, Venkatram Ramaswamy, Steven H. Cohen. 2001. Choice menus for mass customization. *J. Marketing Res.* **38** 183–196.

A multivariate probit model is used to model conjoint data where respondents can select multiple items from a menu. The observed binomial data is modeled with a latent multivariate normal distribution.

Lenk, Peter, Ambar Rao. 1990. New models from old: Forecasting product adoption by hierarchical Bayes procedures. *Marketing Sci.* **9** 42–53.

The nonlinear likelihood function of the Bass model is combined with a random-effects specification across new product introductions. The resulting distribution of heterogeneity is shown to improve early predictions of new product introductions.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, Martin R. Young. 1996. Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Sci.* **15** 173–191.

Fractionated conjoint designs are used to assess ability of the distribution of heterogeneity to "bridge" conjoint analyses across respondents to impute part-worths for attributes not examined.

Manchanda, Puneet, Asim Ansari, Sunil Gupta. 1999. The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Sci.* **18** 95–114.

Multicategory demand data are modeled with a multivariate probit model. Identifying restrictions in the latent error covariance matrix require use of a modified Metropolis-Hastings algorithm.

Marshall, Pablo, Eric T. Bradlow. 2002. A unified approach to conjoint analysis models. *J. Amer. Statist. Assoc.* **97** 674–682.

Various censoring mechanisms are proposed for relating observed interval, ordinal, and nominal data to a latent linear conjoint model.

McCulloch, Robert E., Peter E. Rossi. 1994. An exact likelihood analysis of the multinomial probit model. *J. Econometrics* **64** 217–228.

The multinomial probit model is estimated using data augmentation methods. Approaches to handling identifying model identification are discussed.

Moe, Wendy, Peter Fader. 2002. Using advance purchase orders to track new product sales. *Marketing Sci.* **21** 347–364.

A hierarchical model of product diffusion is developed for forecasting new product sales. The model features a mixture of Weibulls as the basic model, with a distribution of heterogeneity over related products. The model is applied to data on music album sales.

Montgomery, Alan L. 1997. Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Sci.* **16** 315–337.

Bayesian hierarchical models are applied to store-level scanner data. The model specification involves store-level demographic variables. Profit opportunities for store-level pricing are explored using constraints on the change in average price.

Montgomery, Alan L., Eric T. Bradlow. 1999. Why analyst overconfidence about the functional form of demand models can lead to overpricing. *Marketing Sci.* **18** 569–583.

The specification of a function form involves imposing exact restrictions in an analysis. Stochastic restrictions are introduced via a more flexible model specification and prior distribution, resulting in less aggressive policy implications.

Montgomery, Alan L., Peter E. Rossi. 1999. Estimating price elasticities with theory-based priors. *J. Marketing Res.* **36** 413–423.

The prior distribution is used to stochastically impose restrictions on price elasticity parameters that are consistent with economic theory. This proposed approach is compared to standard shrinkage estimators that employ the distribution of heterogeneity.

Neelamegham, Ramya, Pradeep Chintagunta. 1999. A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Sci.* **18** 115–136.

Alternative information sets are explored for making new product forecasts in domestic and international markets, using a Poisson model for attendance with log-normal heterogeneity.

Putler, Daniel S., Kirthi Kalyanam, James S. Hodges. 1996. A Bayesian approach for estimating target market potential with limited geodemographic information. *J. Marketing Res.* **33** 134–149.

Prior information about correlation among variables is combined with data on the marginal distribution to yield a joint posterior distribution.

Rossi, Peter E., Zvi Gilula, Greg M. Allenby. 2001. Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *J. Amer. Statist. Assoc.* **96** 20–31.

Consumer response data on a fixed-point rating scale are assumed to be censored outcomes from a latent normal distribution. Variation in the censoring cutoffs among respondents allow for scale use heterogeneity.

Rossi, Peter E., Robert E. McCulloch, Greg M. Allenby. 1996. The value of purchase history data in target marketing. *Marketing Sci.* **15** 321–340.

The information content of alternative data sources is evaluated using an economic loss function of coupon profitability. The value of a household's purchase history is shown to be large relative to demographic information and other information sets.

Rossi, Peter E., Greg M. Allenby. 1993. A Bayesian approach to estimating household parameters. *J. Marketing Res.* **30** 171–182.

Individual-level parameters are obtained with the use of an informative, but relatively diffuse, prior distribution. Methods of assessing and specifying the amount of prior information are proposed.

Sandor, Zsolt, Michel Wedel. 2001. Designing conjoint choice experiments using managers' prior beliefs. *J. Marketing Res.* **28** 430–444.

The information from an experiment involving discrete choice models depends on the experimental design and the values of the model parameters. Optimal designs are determined with an information measure that is dependent on the prior distribution.

Seetharaman, P. B., Andrew Ainslie, Pradeep Chintagunta. 1999. Investigating household state dependence effects across categories. *J. Marketing Res.* **36** 488–500.

Multiple scanner panel datasets are used to estimate a model of brand choice with state dependence. Individual-level estimates of state dependence effects are examined among categories.

Shively, Thomas A., Greg M. Allenby, Robert Kohn. 2000. A nonparametric approach to identifying latent relationships in hierarchical models. *Marketing Sci.* **19** 149–162.

Stochastic splines are used to explore the covariate specification in the distribution of heterogeneity. Evidence of highly nonlinear relationships is provided.

Steenburgh, Thomas J., Andrew Ainslie, Peder H. Engebretson. 2002. Massively categorical variables: Revealing the information in zipcodes. *Marketing Sci.* **22** 40–57.

The effects associated with massively categorical variables, such as zip codes, are modeled in a random-effects specification. Alternative loss functions are examined for assessing the value of the resulting shrinkage estimates.

Talukdar, Debabrata, K. Sudhir, Andrew Ainslie. 2002. Investing new production diffusion across products and countries. *Marketing Sci.* **21** 97–116.

The Bass diffusion model is coupled with a random effects specification for the coefficients of innovation, imitation, and market potential. The random effects model includes macroeconomic covariates that have large explanatory power relative to unobserved heterogeneity.

Ter Hofstede, Frenkel, Michel Wedel, Jan-Benedict E. M. Steenkamp. 2002. Identifying spatial segments in international markets. *Marketing Sci.* **21** 160–177.

The distribution of heterogeneity in a linear regression model is specified as a conditional Guassian field to reflect spatial associations. The heterogeneity specification avoids the assumption that the random effects are globally independent.

Ter Hofstede, Frenkel, Youingchan Kim, Michel Wedel. 2002. Bayesian prediction in hybrid conjoint analysis. *J. Marketing Res.* **34** 253–261.

Self-state attribute-level importance and profile evaluations are modeled as joint outcomes from a common set of partworths. The likelihoods for the dataset differ and include other, incidental parameters that facilitate the integration of information to produce improved estimates.

Wedel, Michel, Rik Pieters. 2000. Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Sci.* **19** 297–312.

A multilevel model of attention and memory response is used to investigate the effect of brand, pictorial, and text attributes of print advertisements. Information in the data is integrated through a multilayered likelihood specification.

Yang, Sha, Greg M. Allenby. 2000. A model for observation, structural, and household heterogeneity in panel data. *Marketing Lett.* **11** 137–149.

Structural heterogeneity is specified as a finite mixture of nonnested likelihoods, and covariates are associated with the mixture point masses.

Yang, Sha, Greg M. Allenby, Geraldine Fennell. 2002a. Modeling variation in brand preference: The roles of objective environment and motivating conditions. *Marketing Sci.* **21** 14–31.

Intraindividual variation in brand preference is documented and associated with variation in the consumption context and motivations for using the offering. The unit of analysis is shown be at the level of a person-occasion, not the person.

Yang, Sha, Greg M. Allenby. 2003. Modeling interdependent consumer preferences. *J. Marketing Res.* Forthcoming.

The distribution of heterogeneity is modeled using a spatial autoregressive process, yielding interdependent draws from the mixing distribution. Heterogeneity is related to multiple networks defined with geographic and demographic variables.

References

Ainslie, Andrew, Peter Rossi. 1998. Similarities in choice behavior across product categories. *Marketing Sci.* **17** 91–106.

Allenby, Greg M., Neeraj Arora, James L. Ginter. 1995. Incorporating prior knowledge into the analysis of conjoint studies. *J. Marketing Res.* **32** 152–162.

——, ——, ——. 1998. On the heterogeneity of demand. *J. Marketing Res.* **35** 384–389.

——, James L. Ginter. 1995. Using extremes to design products and segment markets. *J. Marketing Res.* **32** 392–403.

——, Peter J. Lenk. 1994. Modeling household purchase behavior with logistic normal regression. *J. Amer. Statist. Assoc.* **89** 1218–1231.

——, ——. 1995. Reassessing brand loyalty, price sensitivity, and merchandising effects on consumer brand choice. *J. Bus. Econom. Statist.* **13** 281–289.

——, Robert P. Leone, Lichung Jen. 1999. A dynamic model of purchase timing with application to direct marketing. *J. Amer. Statist. Assoc.* **94** 365–374.

——, Peter E. Rossi. 1999. Marketing models of consumer heterogeneity. *J. Econometrics* **89** 57–78.

Andrews, Rick, Asim Ansari, Imran Currim. 2002. Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *J. Marketing Res.* 87–98.

Ansari, Asim., Skander Essegaier, Rajeev Kohli. 2000a. Internet recommendation systems. *J. Marketing Res.* **37** 363–375.

——, Kamel Jedidi, Sharan Jagpal. 2000b. A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Sci.* **19** 328–347.

Arora, Neeraj, Greg M. Allenby. 1999. Measuring the influence of individual preference structures in group decision making. *J. Marketing Res.* **36** 476–487.

——, Greg M. Allenby, James L. Ginter. 1998. A hierarchical Bayes model of primary and secondary demand. *Marketing Sci.* **17** 29–44.

Barnard, John, Robert E. McCulloch, Xiao-Li Meng. 2000. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10** 4–24.

Bernardo, Jose, Adrian F. M. Smith. 1994. *Bayesian Theory*. John Wiley, New York.

Blattberg, Robert C., Edward I. George. 1991. Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *J. Amer. Statist. Assoc.* **86** 304–315.

Boatwright, Peter, Robert McCulloch, Peter E. Rossi. 1999. Account-level modeling for trade promotion: An application of a constrained parameter hierarchical model. *J. Amer. Statist. Assoc.* **94** 1063–1073.

Bradlow, Eric T., S. Fader. 2001. A Bayesian lifetime model for the "Hot 100" *Billboard* songs. *J. Amer. Statist. Assoc.* **96** 368–381.

——, Vithala R. Rao. 2000. A hierarchical Bayes model for assortment choice. *J. Marketing Res.* **37** 259–268.

Chang, Kwangpil, S. Siddarth, Charles B. Weinberg. 1999. The impact of heterogeneity in purchase timing and price responsiveness on estimates of sticker shock effects. *Marketing Sci.* **18** 178–192.

Chiang, Jeongwen, Siddartha Chib, Chakravarthi Narasimhan. 1999. Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. *J. Econometrics* **89** 223–248.

Chib, Siddartha. 2003. Monte Carlo methods and Bayesian computation: Overview. S. E. Fienberg, J. B. Kadane, eds. *International Encyclopedia of the Social and Behavioral Sciences: Statistics*. Elsevier Science, Amsterdam, The Netherlands. In press.

DeSarbo, Wayne, Youngchan Kim, Duncan Fong. 1999. A Bayesian multidimensional scaling procedure for the spatial analysis of revealed choice data. *J. Econometrics* **89** 79–108.

Dube, Jean-Pierre. 2003. Multiple discreteness and product differentiation: Demand for carbonated soft drinks. *Marketing Sci.* Forthcoming.

Edwards, Yancy, Greg M. Allenby. 2002. Multivariate analysis of multiple response data. *J. Marketing Res.* Forthcoming.

Efron, Brad, Carl Morris. 1975. Data analysis using Stein's estimator and its generalizations, *J. Amer. Statist. Assoc.* **70** 311–319.

Erdem, Tulin, Micheal Keane. 2003. Brand and quantity choice dynamics under price uncertainty. *Quantitative Marketing Econom.* **1** 5–64.

Frühwirth-Schnatter, Sylvia, Regina Tückler, Thomas Otter. 2003. Bayesian analysis of the heterogeneity model. *J. Bus. Econom. Statist.* Forthcoming.

Gelfand, Alan E., Adrian F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **87**(June) 523–532.

Gelman, Andrew, John B. Carlin, Hal S. Stern, Donald B. Rubin. 1995. *Bayesian Data Analysis*. Chapman Hall, London.

Gilbride, Tim, Greg Allenby. 2003. Attribute-based consideration sets. Working paper, Ohio State University.

Heckman, James, Bernard Singer. 1982. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **5 2** 271–320.

Huber, Joel, Kenneth Train. 2001. On the similiarity of classical and Bayesian estimates of individual mean partworths. *Marketing Lett.* **12** 259–269.

Jen, Lichung, Chien-Heng Chou, Greg M. Allenby. 2003. A Bayesian approach to modeling purchase frequency. *Marketing Lett.* **14** 5–20.

Kalyanam, Kirthi. 1996. Pricing decision under demand uncertainty: A Bayesian mixture model approach. *Marketing Sci.* **15** 207–221.

——, Thomas S. Shively. 1998. Estimating irregular pricing effects: A stochastic spline regression approach. *J. Marketing Res.* **35** 16–29.

Kamakura, Wagner A., Michel Wedel. 1997. Statistical data fusion for cross-tabulation. *J. Marketing Res.* **34** 485–498.

Keane, Michael. 1993. Simulation estimation methods for limited dependent variable models. G. S. Maddala, C. R. Rao, H. D. Vinod, eds. *Handbook of Statistics*, Vol. 11. North Holland, Amsterdam, The Netherlands.

Kim, Jaehwan, Greg M. Allenby, Peter E. Rossi. 2002. Modeling consumer demand for variety. *Marketing Sci.* **21** 223–228.

Leichty, John, Venkatram Ramaswamy, Steven H. Cohen. 2001. Choice menus for mass customization. *J. Marketing Res.* **38** 183–196.

Lenk, Peter J., Wayne S. DeSarbo 2000. Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika* **65** 93–119.

——, Ambar Rao. 1990. New models from old: Forecasting product adoption by hierarchical Bayes procedures. *Marketing Sci.* **9** 42–53.

——, Wayne S. DeSarbo, Paul E. Green, Martin R. Young. 1996. Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Sci.* **15** 173–191.

Liu, Jun S. 2001. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

Manchanda, Puneet, Asim Ansari, Sunil Gupta. 1999. The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Sci.* **18** 95–114.

——, Pradeep K. Chintagunta, Peter E. Rossi. 2003. Response modeling with non-random marketing mix variables. Working paper, Graduate School of Business, University of Chicago, Chicago, IL.

Marshall, Pablo, Eric T. Bradlow. 2002. A unified approach to conjoint analysis models. *J. Amer. Statist. Assoc.* **97** 674–682.

McCulloch, Robert, Nicholas Polson, Peter Rossi. 2000. Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99** 173–193.

——, Peter E. Rossi. 1994. An exact likelihood analysis of the multinomial probit model. *J. Econometrics* **64** 217–228.

——, ——. 1999. Bayesian analysis of multinomial probit model. Mariano, Weeks, Schuermann, eds. *Simulation-Based Inference in Econometrics*. Cambridge University, Cambridge, U.K.

McFadden, Daniel. 2001. Economic choices. *Amer. Econom. Rev.* **91** 351–370.

Mehta, Nitin, Kannan Srinivasan. 2003. Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing Sci.* **22** 58–84.

Montgomery, Alan L. 1997. Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Sci.* **16** 315–337.

——, Eric T. Bradlow. 1999. Why analyst overconfidence about the functional form of demand models can lead to overpricing. *Marketing Sci.* **18** 569–583.

——, Peter E. Rossi. 1999. Estimating price elasticities with theory-based priors. *J. Marketing Res.* **36** 413–423.

Morris, Carl. 1983. Parametric empirical bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65.

Neelamegham, Ramya, Pradeep Chintagunta. 1999. A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Sci.* **18** 115–136.

Newton, Michael, Adrian E. Raftery. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Royal Statist. Soc. Series B* **56** 3–48.

Nobile, Augustino. 1998. A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statist. Comput.* **8** 229–242.

Otter, Thomas, Sylvia Frühwirth-Schnatter, Regina Tüchler. 2003. Unobserved preference changes in conjoint analysis. Working paper, University of Vienna.

Putler, Daniel S., Kirthi Kalyanam, James S. Hodges. 1996. A Bayesian approach for estimating target market potential with limited geodemographic information. *J. Marketing Res.* **33** 134–149.

Robert, Christian P., George Casella. 1999. *Monte Carlo Statistical Methods*. Springer, New York.

Rossi, Peter E., Greg M. Allenby. 1993. A Bayesian approach to estimating household parameters. *J. Marketing Res.* **30** 171–182.

——, Zvi Gilula, Greg M. Allenby. 2001. Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *J. Amer. Statist. Assoc.* **96** 20–31.

——, Robert E. McCulloch, Greg M. Allenby. 1996. The value of purchase history data in target marketing. *Marketing Sci.* **15** 321–340.

Sandor, Zsolt, Michel Wedel. 2001. Designing conjoint choice experiments using managers' prior beliefs. *J. Marketing Res.* **28** 430–444.

Sawtooth Software. 2001. CBC hierarchical Bayes analysis technical paper. Sawtooth Software Technical Paper Series, *www.sawtoothsoftware.com*.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

Seetharaman, P. B., Andrew Ainslie, Pradeep Chintagunta. 1999. Investigating household state dependence effects across categories. *J. Marketing Res.* **36** 488–500.

Shively, Thomas A., Greg M. Allenby, Robert Kohn. 2000. A nonparametric approach to identifying latent relationships in hierarchical models. *Marketing Sci.* **19** 149–162.

Steenburgh, Thomas J., Andrew Ainslie, Peder H. Engebretson. 2002. Massively categorical variables: Revealing the information in zip codes. *Marketing Sci.* **22** 40–57.

Tanner, Martin A., Wing H. Wong. 1987. The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550.

Ter Hofstede, Frenkel, Michel Wedel, Jan-Benedict E. M. Steenkamp. 2002. Identifying spatial segments in international markets. *Marketing Sci.* **21** 160–177.

——, Youingchan Kim, Michel Wedel. 2002. Bayesian prediction in hybrid conjoint analysis. *J. Marketing Res.* **34** 253–261.

Tierney, Luke. 1994. Markov chains for exploring posterior distributions. *Ann. Statist.* **23**(4) 1701–1728.

Wedel, Michel, Rik Pieters. 2000. Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Sci.* **19** 297–312.

Yang, Sha, Greg M. Allenby. 2000. A model for observation, structural, and household heterogeneity in panel data. *Marketing Lett.* **11** 137–149.

——, ——. 2003. Modeling interdependent consumer preferences. *J. Marketing Res.* Forthcoming.

——, ——, Geraldine Fennell. 2002. Modeling variation in brand preference: The roles of objective environment and motivating conditions. *Marketing Sci.* **21** 14–31.

——, Yuxin Chen, Greg Allenby. 2003. Bayesian analysis of simultaneous demand and supply. Working paper, Ohio State University.

This paper was received July 28, 2002, and was with the authors 2 months for 2 revisions; processed by Pradeep Chintagunta.