Essays on Customization Applications in Marketing

Adiguzel, Feray

Publication date:
2006

Chapter 2
Decision Making under Uncertainty

2.1 Introduction

Bayesian decision analysis is a powerful tool to many professionals: market researchers, operations researchers, statisticians, businessmen, economists, engineers, psychologists, computer scientists and those in other fields where prediction and decision making must follow from statistical analysis. The Bayesian statistical tradition provides a formalized way of learning about the parameters of a statistical model from data and originated in 1763, with the theorem formulated by Reverend Thomas Bayes. The Bayesian paradigm has received tremendous popularity in marketing, since it affords an exceedingly flexible and robust framework for developing and estimating statistical models that facilitate realistic description of marketing data. Such models may include latent variables, missing data, mixed outcome data, heterogeneity of coefficients, and more. In particular, in Chapter 3 the model formulated for survey data includes missing values (due to the design of questionnaires) and possibly mixed outcome variables (i.e. rating scales, binary pick-any items, categorical demographic variables, etc.). In Chapter 4, we develop a model with heterogeneity of coefficients (individual level price and promotion sensitivities) and mixed outcomes (the incidence of a category and the expenditure on it). A basic paradigm in marketing is the notion that customers differ in their preferences, needs and choices, and that firms need to take such differences into account in determining optimal marketing actions. Rossi and Allenby (2003) postulate that statistical analysis of marketing data is comprised of three components: within-unit behavior and across unit heterogeneity in that behavior (where unit can be a consumer, a household, or an organization), and action -the solution to the marketing decision problem that recognizes these previous components. Marketing data typically is comprised of many heterogeneous units, often with only limited information on each unit. The statistical problems associated with accommodating heterogeneity of consumers in statistical models and the subsequent management decisions have propelled the use of Bayesian statistical methodology.

It is fair to say that in marketing, the Bayesian paradigm is now the dominant paradigm for inference and decision making in such diverse areas as pricing, new product development, promotions, conjoint analysis and the design of conjoint experiments and --of particular interest to this thesis-- customization of marketing instruments to individual consumers. Some applications of Bayesian approach are in pricing (Montgomery, 1997, Montgomery et al., 1999, Kalyanam et al., 1998, Kalyanam, 1996); in new product development (Neelamegram et al., 1999, Lenk et al., 1990, Allenby et al., 1995, Talukdar et al., 2002, Michalek et al. 2005); in promotions (Blattberg et al., 1991, Boatwright et al., 1999); in conjoint analysis (Allenby et al., 1995, Andrews et al., 2002, Marshall et al., 2002, Otter et al., 2003, Bradlow et al., 2004); design of conjoint experiments (Sandor et al., 2001, Lenk et al., 1996); consumer demand modeling (Allenby et al., 1998, Kim et al., 2002); advertising (Wedel et al., 2000); in customization (Ansari et al., 2003, Liechty et al., 2001); and Internet applications, such as recommendation engines, web-browsing behavior etc. (Ansari et al., 2000, Bradlow et al., 2000, Sismeiro et al., 2004, Rahul et al., 2004). See also the review of Rossi and Allenby (2003).

Rossi, McCulloch and Allenby state in their seminal paper in 1996: “Any successful customization approach must deal directly with the problem of partial information and take parameter uncertainty into account in the decision problem”. In this thesis, we will apply the Bayesian paradigm to facilitate decisions on “what to ask whom” in the construction split questionnaires and on “what to promote to whom” in designing optimal promotional plans, in Chapters 3 and 4 of this thesis. In this chapter, we discuss the Bayesian paradigm, advantages of this approach on a classical approach, and Bayesian decision making that will be used commonly later. In this chapter, we follow Rossi and Allenby (2003), Rossi, McCulloch and Allenby (2005) and Lenk and Wedel (2001).

Bayesian methods propose the optimal way to make consistent decisions in the face of uncertainty. The reason behind this is that Bayesian statistics seek to optimally combine information from two sources: the information that we have or believe at the start of the research and the information in the observed data. Bayes theorem provides the mechanism to combine these both sources of information into a single set of updated information (i.e. the posterior distribution) of the quantities of interest.

In statistics, we quantify uncertainty in observable scientific data through probability distributions, which depend on unknown quantities, called parameters. In the Bayesian paradigm, current knowledge before data analysis is represented with a prior distribution, and updated knowledge based on available data is represented with a posterior distribution. Explicitly, current knowledge about the model parameters is expressed by placing a probability distribution on the parameters, called the "prior distribution", often written as $p(\theta)$. This prior distribution can take on a variety of well known forms, such as the Normal, Binomial, Bernoulli, Poisson, and Gamma distributions, but also multivariate distributions such as the Multivariate Normal, Multinomial, Dirichlet and Wishart (see Casella and Berger, 1990, for an overview of these distributions and specific details). When new data *y* becomes available, the information they contain regarding the model parameters is expressed in the "likelihood," which is proportional to the distribution of the observed data given the model parameters, and written as $p(y \mid \theta)$. Thus, the likelihood requires the specification of a distributional form for the data, as a function of the unknown parameters θ. The information in the data as contained in the likelihood is then combined with the prior to produce an updated probability distribution called the "posterior distribution," $p(\theta, y)$ on which all Bayesian inference is based. Bayes' theorem illustrates how this update is done mathematically and shows that the posterior is proportional to the prior times the likelihood,

$$p(\theta \mid y) = \frac{p(\theta) \times p(y \mid \theta)}{\int p(\theta) \times p(y \mid \theta) d\theta} \propto p(\theta) \times p(y \mid \theta) \qquad (2.1)$$

This posterior distribution captures the uncertainty in the parameters after the data has been observed, but can take complex forms for realistic models. A major breakthrough in the application of the Bayesian paradigm was the realization that in many cases the expressions for the joint posterior distribution of multiple parameters can be factored into simpler expressions that can be recursively sampled from, using so called Markov Chain Monte Carlo methods (MCMC, Geman & Geman 1984, Gelfand & Smith 1990). Markov Chain Monte Carlo (MCMC) simulation has enabled the estimation of complex models that are nearly impossible or very difficult to estimate with classical methods (Gelfand and Smith 1990, Smith and Roberts 1993, Gilks et al. 1996). This holds in particular for hierarchical Bayes models, which have received much popularity in marketing because of the importance of individual differences in a wide variety of models, and the need to investigate factors that influence those. A main theoretical advantage of the Bayesian framework is that while we examine the probability of the data given a model (hypothesis) in frequentist statistics, we examine the posterior probability of a model --or its parameters-- given the data in Bayesian statistics. This, for example, allows for accurate inference in small samples.

In Markov Chain Monte Carlo algorithms, one generates a large sample of independent draws from the posterior distribution, and each draw is conditional on the previous one. MCMC is a Monte Carlo integration using Markov chains. The transition probabilities between sample values are only a function of the most recent value, which is why the technique is referred to as a Markov Chain. The Monte Carlo term comes from the Monte Carlo integration in which we draw samples from the distribution, and then calculate sample averages to approximate expectations. The Gibbs sampler (Geman and Geman 1984) is the most commonly used Markov

Chain Monte Carlo method and widely applicable to various Bayesian problems. We used the Gibbs sampler to impute missing data in Chapter 3 based on a survey response model, and to estimate the parameters of a hierarchical Bayes multivariate type-2 tobit model in Chapter 4.

Gibbs sampling simply means sampling from the full conditional distributions. Suppose that there is a random vector Y which consists of J subvectors, $Y=(Y_1,Y_2,\ldots,Y_J)$, and the joint distribution of Y, P(Y). We iteratively draw from the conditional distribution of each subvector given all the others in Gibbs sampling. We represent this given the value of Y at each step t,

$$\begin{aligned} Y_1^{(t+1)} &\sim P(Y_1 \mid Y_2^{(t)}, Y_3^{(t)}, \ldots\ldots, Y_J^{(t)}) \\ Y_2^{(t+1)} &\sim P(Y_2 \mid Y_1^{(t+1)}, Y_3^{(t)}, \ldots\ldots, Y_J^{(t)}) \qquad (2.2) \\ &. \\ Y_J^{(t+1)} &\sim P(Y_J \mid Y_1^{(t+1)}, Y_2^{(t+1)}, \ldots\ldots, Y_{J-1}^{(t+1)}) \end{aligned}$$

Gibbs sampling is closely related to data augmentation (Tanner and Wong, 1987). Data augmentation refers to methods for constructing iterative algorithms via introduction of unobserved data or latent variables. Many models in marketing which contain latent variables use this method, including limited dependent variable models (like choice or censored regression models), state space, or common factor models, and models with heterogeneity. Gibbs sampling is especially well suited to coping with incomplete information and we illustrated the application of Gibbs sampling on missing data in Chapter 3.

Loss functions are used to estimate parameters in the Bayesian approach and to make decisions. A loss function measures the loss caused by an estimation error or decision error. Estimation is a special case in decision-making, and the goal is to choose the estimator which minimizes the expected loss. In the case of estimation problems, the loss function is a function of the parameter estimate and the true (unknown) parameter value. Common choices of loss functions are quadratic loss, absolute loss and zero-one loss, which are represented below, respectively.

$$L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2 \tag{2.3}$$

$$L(\hat{\theta},\theta) = \left|\hat{\theta}-\theta\right| \tag{2.4}$$

$$L(\hat{\theta},\theta) \begin{array}{lll} = 0 & \text{if} & \hat{\theta}=\theta \\ = c & \text{if} & \hat{\theta}=\theta \end{array} \tag{2.5}$$

While the optimal Bayesian point estimate is the mean of the posterior distribution when choosing a quadratic loss function, the median of the posterior distribution is the point estimate when choosing the absolute loss function. The mode of the posterior distribution is the point estimate in the case of a zero-one loss function. Although the choice of loss function depends on the problem, the most commonly used is the quadratic loss function, which uses all the information in the posterior distribution. We will focus on the case of decision problems later.

We briefly discuss a few issues of the Bayesian approach, in particular the choice of the prior distribution. There are two kinds of prior information: objective and subjective (see more information on objective and subjective priors in Press, 2003). In the classical approach, all modeling assumptions are actually a kind of prior information, and generally the underlying assumptions can remain hidden, whereas the researcher’s prior beliefs are expressed as prior information in the Bayesian approach. Prior beliefs can be obtained from experts, prior studies, theories or other data sets. Prior distributions are intrinsically subjective (everyone’s prior information is different). Subjective Bayesian statistics, firmly rooted in probability theory (De Finetti, 1970), proposes that a model reflects a researcher's belief about a phenomenon and that people can and should conceive of uncertainty about events as subjective probabilities (Savage, 1954). This at the same time raises a reservation some have about the Bayesian approach: posterior predictive inferences are sensitive to the choice of the prior, and so are decisions made based on the model inferences. Many “pragmatic Bayesians” (See Lenk and Wedel, 2001), predominantly concerned with the flexibility of model construction that Bayesian statistics now afford through MCMC methodology, therefore choose non-informative priors for their model. Many statistical models formulated in the Bayesian framework, are therefore based on non-informative or weakly informative priors, which minimize the influence of prior assumptions on posterior inference. Although many statisticians see the subjective priors as a fundamental drawback of the Bayesian approach, this is inescapable, and frequentist methods (classical statistical theory) also entail subjective choices. However, using prior information well may in fact improve decision-making (Berger 1985). Certainly, the investigation of the sensitivity of the predictive distribution to the specification of the prior is critical. In case subjective prior distributions for the model parameters can be assessed, we may need to elicit priors from consumers, decision makers or other subject-matter experts. Allenby et al. (1995) state that incorporating expected prior ordinal information of attributes into conjoint analysis improves the estimation. Sandor and Wedel (2001) illustrate that this approach is attractive in the design of choice experiments. Popkowski and Sinha (2005) propose a method to facilitate the subjective information from a modeler or manager into a choice model and illustrate the improvement on the marketing strategy (decisions). Wolfson (1995) and Chaloner (1996) provide an overview of the various philosophies of elicitation based on the ways people think about and update probabilistic statements. The use of loss functions to make optimal decisions concerning settings of control variables allows the statistical process to be customized to fit the particular application in question. Emphasis in Bayesian marketing is now shifting from inference towards the decision problem. Although quickly improving in quality, the models may not realize their full potential in decision making until put into the framework of the decision-making process. Better decisions require better procedures to extract information from data and incorporate that in the marketing decision problem, and the Bayesian paradigm is optimally suited for that (Lenk and Wedel, 2001). I provide two examples of this in Chapters 3 and 4 of this thesis.

2.2 Bayesian Analysis and Marketing Decisions

In the previous section, we discussed the Bayesian approach for statistical inference and here we discuss the Bayesian approach for decision-making. The main purpose of decision theory is to develop techniques and methods that facilitate making decisions in an optimal way. The standard estimation problems of statistical inference or testing hypotheses can be formulated as decision problems (see Cyert and DeGroot, 1987). In this section, we compare "Bayesian approach" in decision theory to the naïve approach, "plug-in."

Most of the model parameters on which marketing action decisions are based, are unknown random variables, and we need to choose the optimal value of deterministic control variables, the effects of which depend on those random parameter values. Thus in the decision process, the parameter estimation or model uncertainty must be considered. The Bayesian approach merges all available prior and observed data information to estimate the parameters that are the basis of the optimization problem. Levels for the control variables can then be found that maximize or minimize the expected value of an objective function (Berger, 1985). Bayesian decision theory is very appropriate for marketing decision problems. The reasons are 1) subjective prior probabilities from economic theories, experts or managers can be easily used to express pre-existing information as prior information that improve the estimation, 2) it entails careful modeling of the data structure, checking and allowance for uncertainty in model assumptions, 3) formulating a set of possible decisions can be easy, and 4) utility functions are used to express how the value of each alternative decision is affected by the unknown model parameters. The objective in decision-making is to choose an action that minimizes the expected value of the loss function with respect to the posterior distribution, if data are available. However, if data are not available, the expected loss should be minimized with respect to the prior distribution[2] (see Sandor and Wedel, 2001, on designing conjoint experiments).

The main goal of Bayesian decision theory is to minimize the expected loss of a decision or minimize the expected risk. To do this, Bayesian decision theory leads to an optimal decision considering the expected loss of all possible values of the random parameters values by weighting those by the probability of their occurrence (i.e. posterior distribution). A loss function and the posterior distribution are the two main components in decision theory and will be explained below. In Chapter 3, we develop a model for survey data, with a missing data structure that is due to the design of the questionnaire. Then, we formulate a loss function that captures how far a data-structure with missing data is from a true, complete dataset. The decisions are then what questions to pose to which respondents, to minimize the loss. In Chapter 4, the decision is what categories to promote to which customers, based on a joint model of category incidence and expenditure sensitivity to prices and promotions.

[2] The mean of the posterior distribution is the Bayes estimator with respect to the quadratic loss function. If no data are available, the posterior distribution reduces to the prior distribution, and the Bayes estimator becomes the mean of the prior distribution (see details Press, 2003).

The loss function is related to the incremental revenue obtained from promoting a specific set of categories.

Suppose a manager has two or more alternative actions, “a”, for any decision. Each action “a” has some potential loss that will depend on the parameters, θ, and this relationship is expressed through the loss function, $L(a,\theta)$. For example, a marketing manager has to decide what category to promote for which consumer (Chapter 4). The different possible promotions are the alternative actions. The state of nature is the likely demand for every possible set of promotions, and revenue is the loss function that relates promotion to demand. We can calculate the expected revenue/profit (or loss) for every possible combination of allocation of promotions and demand. Management objectives are specified as a function of model predictions (and/or parameters), for example the predicted revenue arising from promoting a specific category, and the expected consequences of any particular management action (i.e. every possible set of promotions) are calculated by integrating over the uncertainty in both model parameters and model predictions. If we express this mathematically, the optimal decision maker chooses the action so as to minimize expected loss, where the expectation is taken with respect to the posterior distribution (Rossi and Allenby 2003)

$$\min_a E[L(a)] = \int L(a,\theta)p(\theta \mid data)d\theta \qquad (2.6)$$

This integral can be evaluated either numerically or by Monte Carlo integration. The explicit incorporation of the posterior distribution of the random variables (which includes prior information) makes the decision theory approach Bayesian. In decision-making, uncertainty means that the outcome of a decision maker's action is not exactly predictable because of the unknown parameter values or random error terms. The traditional approach only assumes a probability distribution for error terms but not for the unknown parameter values. Such an approach does not permit the choice of an optimal decision, which reflects estimation uncertainty. However, in the Bayesian approach, prior distributions are assigned for unknown values of the all parameters in the decision problem.

Consider the more complex marketing decision problem in which we have explanatory variables x, consisting of some control variables x_c and the remaining explanatory variables, x_f. For example, we may have promotion as a (0/1) control variable, and price as another explanatory variable. We have a probability distribution $p(y|x,\theta)$ which represents how the dependent variable (outcome) is related to explanatory variables. The decision maker wants to choose x_c to maximize the expected profits or revenue where the expectation is taken over the distribution of the outcome variable. In a fully Bayesian decision approach, this expectation must be taken with respect to the posterior distribution of θ and the predictive conditional distribution of $p(y|x_c,x_f)$ and expressed by Rossi and Allenby (2003) as:

$$\begin{aligned} \pi^*(x_c \mid x_f) &= E_\theta[E_{y|\theta}[\pi(y \mid x_c)]] \\ &= E_\theta\left[\int \pi(y \mid x_c)p(y \mid x_c, x_f, \theta)dy\right] \qquad (2.7) \\ &= E_\theta[\overline{\pi}(x_c \mid x_f, \theta)] \end{aligned}$$

Sometimes, a "plug-in" method is used, where the (maximum likelihood or posterior mode) estimates of any unknown random parameters are inserted into the optimization problem and then the decision problem is solved. In other words, the plug-in approach disregards the uncertainty and assumes that parameters are estimated perfectly. However, the certainty equivalence theorem[3] demonstrates that if and only if the posterior distribution of the parameters is normal and the objective function is linear-quadratic in the unknown parameters the plug-in leads to the same solution as a Bayesian approach (Dorfman, 1997). In Chapter 3, we rely on that result when we use a plug-in estimator to compute the optimal questionnaire design, in the case where all measure variables are normal. However, we do not always have symmetric posterior distributions because of prior information or particular distributional assumptions, and the objective function is not always quadratic. Most of the time objective functions in marketing are based on expectations of the explicit function of profits, and these profit functions are generally not linear. In such cases, taking a Bayesian approach to solving for the optimal control will produce a different answer from the standard method. Expressing this mathematically (Rossi and Allenby, 2003):

$$\pi^{*}(x_c) = E_{\theta|y}[\overline{\pi}(x_c \mid \theta] \neq \overline{\pi}(x_c \mid \hat{\theta} = E_{\theta|y}[\theta]) \qquad (2.8)$$

[3] If the loss function is quadratic and the constraining model is linear and stochastic only by additive random disturbances that are independent of the instruments and whose expected values are zero, then the optimal values of the instruments are the same as if there were no uncertainty (Theil, 1954).

This is the approach we take in Chapter 4 when deriving optimal individual level promotion allocations. Bayesian methods are ideal for cases where we have prior information that shapes posterior distributions and for problems where risk is important and the correct objective function is not quadratic, but skewed. For this reason, Bayesian decision theory is often referred to as decision-making under estimation risk, and it relies on incorporating the uncertainty from estimation process into an optimal decision.