NON-CONJUGATE BAYESIAN ANALYSIS USING THE ANALYTIC HIERARCHY PROCESS AND SAMPLING/IMPORTANCE RESAMPLING

Eugene D. Hahn and Cynthia L. Knott
Department of Management Science, School of Business and Public Management
The George Washington University, Washington, DC 20052
genehahn@gwu.edu, cynth@gwu.edu

Abstract: Bayesian analysis proceeds after a prior distribution has been specified. Restricting Bayesian priors to conjugate distributions may yield sub-optimal representations of an expert's prior beliefs. Additionally, there are some cases in which typical techniques for prior elicitation may have some limitations. A non-conjugate approach to Bayesian inference making use of sampling/importance resampling and the Analytic Hierarchy Process redresses these problems.

Introduction

The Bayesian paradigm allows one to incorporate prior information into statistical models for decision-making. Prior information is combined with data using the axioms of probability, yielding probabilistically coherent posterior distributions for parameters of interest. In this paradigm, there are three general types of quantities: the background information existing at the beginning of the study (represented by what is called the prior distribution, $p(\theta)$), the data on the variables of interest (represented by the likelihood, $p(y|\theta)$), and the updated information resulting from the combination of the prior and the likelihood (known as the posterior, $p(\theta|y)$). These quantities are related to one another by the equation:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta} \tag{1}$$

Equation 1 reveals how prior belief and data interact to influence decision-making in the Bayesian paradigm. We begin with our prior beliefs about the outcome of a probabilistic event. Then, the data is obtained and combined with the prior to produce the posterior, the updated representation of our beliefs. Since the posterior represents all the information that is available, it can then be used in subsequent analyses as a new prior. Hence, updating may be performed sequentially, as each piece of data becomes available, or collectively, when all the data to be gathered has been. The posterior will represent updated belief equivalently whether updating occurs sequentially or collectively.

It is worthwhile to note that an infinite number of prior distributions can be specified, including ones that are bizarre or pathological. Asymptotically, the influence of the prior on the posterior disappears. However, since many analyses are conducted with relatively small sample sizes, the construction of a prior should be undertaken carefully. The credence of small sample-size Bayesian analyses is partially predicated on the usage of a prior distribution that the researcher and/or the consumers of research find reasonable. In many cases, a non-informative prior is used to represent prior ignorance about the distribution of a parameter under consideration. However, in other cases we have relevant information and wish to use it in the conduct of a Bayesian analysis. In this situation a method for eliciting prior information and a method for incorporating that information in a Bayesian analysis are needed.

Berger (1985) has noted that the most common technique is to assume the prior density can be represented reasonably well by a member of the family of prior distributions which permit a conjugate analysis to be performed. For example, a Beta(2, 3) prior may seem to be a reasonable approximation to one's prior beliefs for a binomial process. If the conjugate density does indeed provide a reasonable approximation, obtaining the posterior becomes relatively straightforward. However, this approach may not work in all situations. For example, selection of the parameters of the conjugate distribution is frequently done to be concordant with the subjective prior mean and variance. It may be, however, that the person supplying the probabilities is not easily able to articulate a particular moment of the distribution, such as the variance. For example, the problem may be such that a person cannot generate a specific numeric estimate of the variance with reasonable certainty, or it may be that the person has not received enough exposure to statistical concepts to generate a good estimate of the variance. Alternatively, the person may supply a distribution that is not well approximated by a conjugate distribution.

Another method for constructing a prior offered by Berger is the histogram method. The person supplying the prior probabilities is asked to draw a histogram representing his or her prior beliefs. A problem with this method is that tail areas may not be included. For example, there may be a very small probability that the parameter takes on values above and below the range of the histogram. Two rejoinders may be offered, however. First, the person supplying the probabilities is free to continue specifying the histogram until he or she is feels that the distribution has been reasonably well specified. Second, the use of a conjugate prior, while ensuring that tail areas are represented, does not ensure that tail areas are represented reasonably accurately (Berger, 1985). It is probably inappropriate to ask a person whether, between the fourth and fifth standard deviation, a particular normal distribution's decay toward the asymptote reasonably reflects the person's actual prior probabilities. Berger also mentions another caveat associated with the histogram method, i.e., "the prior density so obtained is somewhat difficult to work with" (1985, p. 77). We will show, however, that this difficulty is readily addressed using Monte Carlo methods.

The Analytic Hierarchy Process and Prior Elicitation

As Yager (1979) has noted, the Analytic Hierarchy Process (AHP; Saaty, 1995) can be used to elicit probability distributions. A matrix of ratios can be constructed such that each ratio expresses the perceived likelihood that one event will obtain as compared to another. The unit eigenvector of the matrix is then extracted. Since the elements of the unit eigenvector sum to one, they can be taken as the probabilities of the component events. Hence, the AHP can be used to generate a histogram of prior probabilities in a straightforward manner. A series of intervals, whose relative likelihoods will subsequently be assessed by pairwise comparisons to yield a probability distribution, are inserted into model as objectives.

There are a few advantages to using the AHP to elicit probabilities. First, the person supplying the judgments need only make pairwise comparisons among stimuli. The study of human cognition on pairwise comparative judgment tasks goes back to the birth of psychology as a field (e.g., Thurstone, 1927). Early research in the emerging field then known as psychophysics sought to understand the nuances of comparative judgment; what was not contested, however, is that discriminability of features often seems to be enhanced by the presence of a similar object for comparison. More recently, the judgment and decision-making literature has shown that performance on decision tasks involving probability can be quite mixed. Bolger and Wright (1992) reviewed twenty judgment studies of experts and concluded that experts are variably calibrated when making probability judgments in their areas of expertise. Since the quality of judgments is negatively related to the cognitive processing load (Gilbert, 1989), we can see that the straightforward nature of the pairwise comparison task may facilitate the performance of the person supplying probability judgments. A second advantage of using AHP to elicit probability judgments is that there is no need for the person supplying judgments to explicitly articulate knowledge about the moments of the probability distribution. In contrast, the conjugate method is most profitably employed with explicit quantification of such parameters. Third, the AHP provides information about the inconsistency of the judgment matrix. This information allows for two further possibilities in the case of substantial inconsistency. First, the person supplying judgments may decide to re-evaluate his or her comparisons so that consistency may be improved. Or, second, the person may decide to re-distribute some weight to tail areas, effectively flattening the prior.

Sampling/Importance Resampling

Historically, the evaluation of integrals by numerical methods in a non-conjugate Bayesian analysis could make a noteworthy imposition on the analyst. However, in recent years a number of Monte Carlo techniques have been developed which facilitate these analyses (e.g., Tanner, 1996). Sampling/importance resampling (SIR), introduced by Rubin (1987) is a general Monte Carlo procedure for simulating posterior distributions (see also Albert, 1993, Smith and Gelfand, 1992). Assume there are two probability distributions $g(\theta)$ and $h(\theta)$ that are relatively similar; moreover, $h(\theta)$ is a distribution which is easy to simulate. The problem is to simulate $g(\theta)$ using $h(\theta)$. The SIR process for solving this problem can be summarized in three key steps as follows (Rubin, 1988). Draw a sample $\theta_1, \ldots, \theta_m, j = 1, \ldots m$ from $h(\theta)$. For each member of the sample, calculate its "importance ratio" or sample weight:

$$w_j = \frac{g(\theta_j)}{h(\theta_j)} \qquad (2)$$

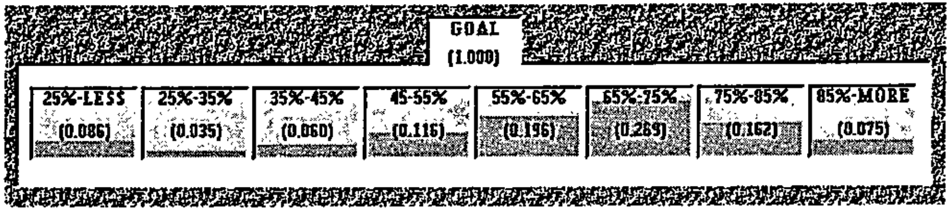
Then take a second sample from $\theta_1, \ldots, \theta_m$, called $\theta'_1, \ldots, \theta'_m$, with probabilities proportional to $w_1, \ldots, w_m$. The density of the second sample, θ'_j, will approximate that of $g(\theta)$. Mapping the relationship between the distributions in a Bayesian analysis and those in the SIR procedure is straightforward. In Equation 1, the desired distribution that is difficult to simulate, $g(\theta)$, is the posterior, $p(\theta|y)$. The distribution $h(\theta)$ is the prior, $p(\theta)$; it is easy to simulate because its distribution function has been specified before commencement of the analysis. The distribution of the ratios, $w(\theta)$, is proportional to $p(y|\theta)$ up to a normalizing constant, c, where c^{-1} is $\int p(\theta)\, p(y|\theta)\, \mathrm{d}\theta$.

The SIR procedure has been described as a "weighted bootstrap" (Smith and Gelfand, 1992). In the typical bootstrap, one samples from θ_j with equal probabilities; here, however, one samples from θ_j with probabilities varying as a function of $g(\theta_j)/h(\theta_j)$. Thus, SIR shares features with p.p.s. sampling methods (Cochran, 1979, ch. 9). Like the bootstrap, the algorithm has quite general applicability, although it may not always be computationally efficient, as we will note momentarily. There are a few caveats associated with the use of SIR to approximate $g(\theta)$. The greater the difference between $g(\theta)$ and $h(\theta)$, the more points, m, should be sampled from θ. In general, the accuracy of the technique increases with m, and decreases with increasing differences between $g(\theta)$ and $h(\theta)$. Moreover, it is important to ensure that the prior is adequately constructed. For example, if the prior is discrete and does not extend beyond a boundary value, b, then the posterior will also not extend beyond b. The area beyond b will occur in the prior with zero probability, and hence will also occur with zero probability in the posterior. For similar reasons, it is often desirable for the tails of the prior to be heavier than those of the posterior. Heavy tails will help to ensure adequate sampling occurs in these regions of relatively low probability. Finally, SIR may provide a poor approximation if the posterior is highly concentrated in a small region of the prior. This is because a resampling of a very small proportion of the prior will generate the posterior. In such an instance, SIR will be an inefficient method of generating the posterior, and m will need to be increased. In general, however, the potential pitfalls of using SIR can be substantially reduced by a little care on the part of the analyst and a willingness to increase m. Moreover, SIR is attractive because it is easy to implement and does not require convergence monitoring as do Markov chain Monte Carlo techniques such as Gibbs sampling.

Bayesian Analysis Using AHP and SIR

We will illustrate the use of AHP in a Bayesian analysis with an example. Suppose that a company recently has made a public offering of its stock. It is of interest to assess whether the stock's price will have gone up or down relative to its current price after 30 days have elapsed. Denote the probability of success (stock price increase) for this binomial event as x. The expert needs to construct a prior over the range of x, where $0 \leq x \leq 1$. In order to create a histogram, the probability space needs to be subdivided into contiguous intervals. The expert feels that the probability of the stock price being up after 30 days is rather likely, but there is a reasonable chance it will fare poorly as well. Moreover, she feels capable of making more refined judgments throughout the range from 25% to 85%. In the tails below 25% and above 85%, however, she feels less capable of making refined judgments. Hence, she decides to have narrower intervals in the range from 25% to 85%, and to have one broader interval for each tail. She decides that the intervals or "bins" for her histogram are: 0% ≤ x < 25%, 25% ≤ x < 35%, 35% ≤ x < 45%, 45% ≤ x < 55%, 55% ≤ x < 65%, 65% ≤ x < 75%, 75% ≤ x < 85%, and 85% ≤ x ≤ 100%. She conducts pairwise comparisons among the alternatives, and arrives at the prior shown in Figure 1. It is of interest to note that this prior would not be well approximated by a Beta distribution, potentially limiting the applicability of conjugate methods.

Figure 1
Prior for Stock Increase



The expert examines the inconsistency index and notes that her judgments do not possess substantial inconsistency. Thus, she decides to proceed with this prior. Note that, in using SIR, the prior is simulated, and then a weighted bootstrap of it is used to generate the posterior. Moreover, we noted that pitfalls can be avoided by monitoring the results of SIR. So, as an initial check on the SIR procedure, we compare the simulated to the actual prior. This check is not intended to be exhaustive, but merely illustrative of possible considerations one might peruse before turning to an examination of the posterior.

Table 1
Means for Actual and Simulated Priors at Different Values of *m*

	Actual Prior	*m = 500*	*m = 1,000*	*m = 5,000*	*m = 10,000*
Mean	0.608125	0.5997219	0.6113059	0.611477	0.609224
Absolute Difference	-	0.0084031	0.0031809	0.003352	0.001099
Percent Error	-	1.40%	0.52%	0.55%	0.18%

Table 1 contains means for the actual prior and the simulated priors with m = 500, 1000, 5000, and 10000. Taking the absolute value of the discrepancy between the simulated prior mean and the actual prior mean as the numerator and the actual prior mean as the denominator, we can examine the percent error associated with values of m for different simulations. We see that at m = 500 the percent error is less than 2%, not terribly large. However, increasing m beyond 500 causes the percent error to drop substantially. In this particular series of simulations, doubling m to equal 1,000 caused percent error to decrease by more than two times, while increasing m twentyfold to equal 10,000 led to a roughly eightfold reduction in percent error. Table 1 also reveals that in these samples percent error did not monotonically decrease with increasing m. These findings serve to underscore the approximate nature of Monte Carlo-based methods. Nonetheless, if we were to conduct an infinite number of replications of each size m, we would expect increased m to result in decreased percent error.

Table 1 suggests that m = 5000 will provide reasonable accuracy without undue computation; hence, we proceed with 5000 as the value for m. After the first month, the stock's price did indeed sell for more than it sold for at the outset of the inquiry. Analysis using SIR generated the posterior shown in Figure 2. The simulated posterior mean after one success and no failures was .684. A 95% credible interval for the mean can be obtained from the .025*m^{th} and .975*m^{th} observations. Here, this interval is (.312 ≤ x ≤ .965).

Over the second month, the stock price increased again, while in the third month it declined. The expert decided to calculate the posteriors sequentially as the data arrived. By doing so, she had the most current information at any given time, for her own use or possibly as an input to another AHP model. Figure 3 shows the prior and posteriors for all three months. The posterior was concentrated in the upper ranges during the second month because of the increase in the price ($\bar{x} = .722$). In the third month, however, the posterior shifted to the left because of the declining price ($\bar{x} = .649$).

Figure 2
Actual Prior and Posterior after One Success

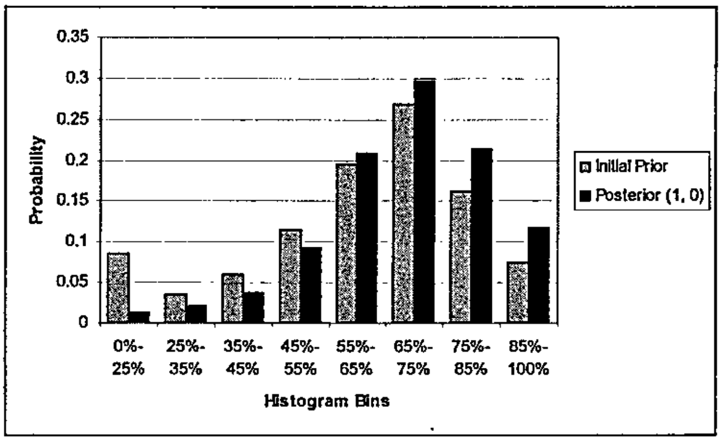
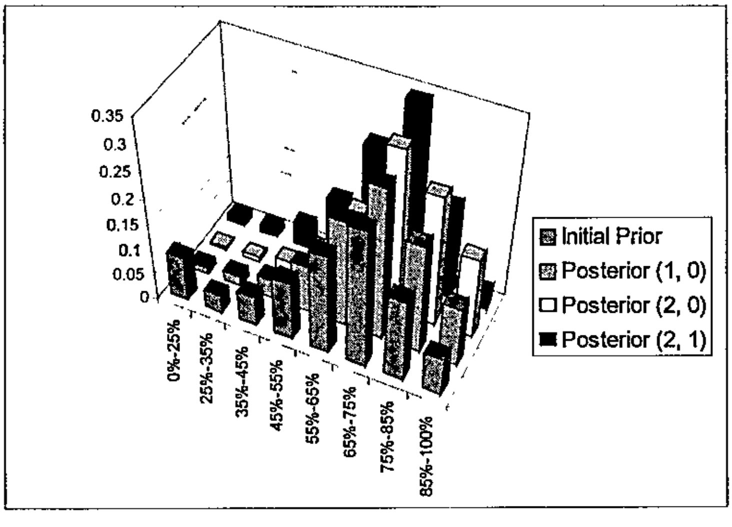


Figure 3
Actual Prior and Sequentially Obtained Posteriors



Conclusion

The construction of a prior that accurately represents prior belief is important in Bayesian analyses, particularly when sample sizes are small. Conjugate priors, while flexible, may not be flexible enough for all cases. In contrast, the histogram-based method can be used to be create a more general class of priors. By using the AHP to create histogram-based priors, experts with or without advanced statistical backgrounds can create priors in a way that capitalizes on the strengths of the cognitive mechanism. SIR can then be used to generate posterior distributions and quantities of interest to any degree of accuracy required. By propagating back and forth the information contained in priors and posteriors, Bayesian analysis can be nicely integrated into a network of AHP models.

References

Albert, J.H. (1993). "Teaching Bayesian Statistics Using Sampling Methods and MINITAB," *The American Statistician*, 47, 182-191.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer.

Bolger, F. and Wright, G. (1992). "Reliability and Validity in Expert Judgments," in *Expertise and Decision Support* (pp. 47-76), eds. G. Wright and F. Bolger, New York: Plenum Press.

Cochran, W.G. (1977). *Sampling Techniques* (3rd. ed.), New York: Wiley.

Gilbert, D.T. (1989). "Thinking Lightly about Others: Automatic Components of the Social Inference Process," in *Unintended Thought* (pp. 189-211), eds. J.S. Uleman and J.A. Bargh, New York: Guilford Press.

Rubin, D.B. (1987). "Comment on 'The Calculation of Posterior Distributions by Data Augmentation,'" by M.A. Tanner and W.H. Wong. *Journal of the American Statistical Association*, 82, 543-546.

Rubin, D.B. (1988). "Using the SIR Algorithm to Simulate Posterior Distributions," in *Bayesian Statistics 3* (pp. 395-402), eds., J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, New York: Oxford University Press.

Saaty, T.L. (1995) *Decision Making for Leaders: The Analytic Hierarchy Process* (3rd ed.), Pittsburgh, PA: RWS Publications.

Smith, A.F.M. and Gelfand, A.E. (1992). "Bayesian Statistics without Tears: A Sampling-Resampling Perspective," *The American Statistician*, 46, 84-88.

Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.), New York: Springer-Verlag.

Thurstone, L.L. (1927). "A Law of Comparative Judgment,"*Psychological Review*, 34, 273-286.

Yager, R.R. (1979). "An Eigenvalue Method of Obtaining Subjective Probabilities," *Behavioral Science*, 24, 382-387.