Bayesian Statistics Without Tears: A Sampling–Resampling Perspective

A. F. M. SMITH and A. E. GELFAND*

Even to the initiated, statistical calculations based on Bayes's Theorem can be daunting because of the numerical integrations required in all but the simplest applications. Moreover, from a teaching perspective, introductions to Bayesian statistics—if they are given at all—are circumscribed by these apparent calculational difficulties. Here we offer a straightforward sampling–resampling perspective on Bayesian inference, which has both pedagogic appeal and suggests easily implemented calculation strategies.

KEY WORDS: Bayesian inference; Exploratory data analysis; Graphical methods; Influence; Posterior distribution; Prediction; Prior distribution; Random variate generation; Sampling–resampling techniques; Sensitivity analysis; Weighted bootstrap.

1. INTRODUCTION

Given data x obtained under a parametric model indexed by finite-dimensional θ, the Bayesian learning process is based on

$$p(\theta|x) = \frac{l(\theta; x)p(\theta)}{\int l(\theta; x)p(\theta)\, d\theta}, \tag{1.1}$$

the familiar form of Bayes's Theorem, relating the posterior distribution $p(\theta|x)$ to the likelihood $l(\theta; x)$, and the prior distribution is $p(\theta)$. If $\theta = (\phi, \psi)$, with interest centering on ϕ, the joint posterior distribution is marginalized to give the posterior distribution for ϕ,

$$p(\phi|x) = \int p(\phi, \psi|x)\, d\psi. \tag{1.2}$$

If summary inferences in the form of posterior expectations are required (e.g., posterior means and variances), these are based on

$$E[m(\theta)|x] = \int m(\theta)p(\theta|x)\, d\theta, \tag{1.3}$$

for suitable choices of $m(\cdot)$.

Thus, in the continuous case, the integration operation plays a fundamental role in Bayesian statistics, whether it is for calculating the normalizing constant in (1.1), the marginal distribution in (1.2), or the expectation in (1.3). However, except in simple cases, explicit evaluation of such integrals will rarely be possible, and realistic choices of likelihood and prior will necessitate the use of sophisticated numerical integration or analytic approximation techniques (see, for example, Smith et al. 1985, 1987; Tierney and Kadane, 1986). This can pose problems for the applied practitioner seeking routine, easily implemented procedures. For the student, who may already be puzzled and discomforted by the intrusion of too much calculus into what ought surely to be a simple, intuitive, statistical learning process, this can be totally off-putting.

In the following sections, we address this problem by taking a new look at Bayes's Theorem from a sampling–resampling perspective. This will open the way to both easily implemented calculations and essentially calculus-free insight into the mechanics and uses of Bayes's Theorem.

2. FROM DENSITIES TO SAMPLES

As a first step, we note the essential duality between a sample and the density (distribution) from which it is generated. Clearly, the density generates the sample; conversely, given a sample we can approximately recreate the density (as a histogram, a kernel density estimate, an empirical cdf, or whatever).

Suppose we now shift the focus in (1.1) from densities to samples. In terms of densities, the inference process is encapsulated in the updating of the prior density $p(\theta)$ to the posterior density $p(\theta|x)$ through the medium of the likelihood function $l(\theta; x)$. Shifting to samples, this corresponds to the updating of a sample from $p(\theta)$ to a sample from $p(\theta|x)$ through the likelihood function $l(\theta; x)$.

In Section 3, we examine two resampling ideas that provide techniques whereby samples from one distribution may be modified to form samples from another distribution. In Section 4, we illustrate how these ideas may be utilized to modify prior samples to posterior samples, as well as to modify posterior samples arising under one model specification to posterior samples arising under another. An illustrative example is provided in Section 5.

3. TWO RESAMPLING METHODS

Suppose that a sample of random variates is easily generated, or has already been generated, from a continuous density $g(\theta)$, but that what is really required is a sample from a density $h(\theta)$ absolutely continuous with

*A. F. M. Smith is Professor, Department of Mathematics, Imperial College of Science Technology and Medicine, London SW7 2BZ, England. A. E. Gelfand is Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269. The authors are grateful to David Stephens for assistance with computer experiments. His work and a visit to the United Kingdom by the second author were supported by the U.K. Science and Engineering Council Complex Stochastic Systems Initiative.

respect to $g(\theta)$. Can we somehow utilize the sample from $g(\theta)$ to form a sample from $h(\theta)$? Slightly more generally, given a positive function $f(\theta)$ which is normalizable to such a density $h(\theta) = f(\theta)/\int f(\theta)\, d\theta$, can we form a sample from the latter given only a sample from $g(\theta)$ and the functional form of $f(\theta)$?

3.1 Random Variates via the Rejection Method

In the case where there exists an identifiable constant $M > 0$ such that $f(\theta)/g(\theta) \leq M$, for all θ, the answer is yes to both questions, and the procedure is as follows:

1. Generate θ from $g(\theta)$.
2. Generate u from uniform (0, 1).
3. If $u \leq f(\theta)/Mg(\theta)$, accept θ; otherwise, repeat Steps 1–3.

Any accepted θ is then a random variate from

$$h(\theta) = f(\theta) \Big/ \int f(\theta)\, d\theta.$$

The proof (see also Ripley 1986, p. 60) is straightforward.

Let

$$S_0 = [(\theta, u): \theta \leq \theta_0, u \leq f(\theta)/Mg(\theta)],$$

and let

$$S = [(\theta, u): u \leq f(\theta)/Mg(\theta)].$$

Then the cdf of accepted θ, according to the preceding procedure, is

$$\begin{aligned}\Pr(\theta \leq \theta_0 | \theta \text{ accepted}) &= \frac{\Pr(\theta \leq \theta_0,\ \theta \text{ accepted})}{\Pr(\theta \text{ accepted})} \\ &= \frac{\iint 1_{S_0} \cdot g(\theta)\, du\, d\theta}{\iint 1_S \cdot g(\theta)\, du\, d\theta} \\ &= \frac{\int_{-\infty}^{\theta_0} f(\theta)\, d\theta}{\int_{-\infty}^{\infty} f(\theta)\, d\theta}.\end{aligned}$$

It follows that accepted θ have density $h(\theta) \propto f(\theta)$.

Hence, for a sample θ_i, $i = 1, \ldots, n$, from $g(\theta)$, in resampling to obtain a sample from $h(\theta)$ we will tend to retain those θ_i for which the ratio of f relative to g is large, in agreement with intuition. The resulting sample size is, of course, random, and the probability that an individual item is accepted is given by

$$\begin{aligned}\Pr(\theta \text{ accepted}) &= \iint 1_S \cdot g(\theta)\, du\, d\theta \\ &= M^{-1} \int_{-\infty}^{\infty} f(x)\, dx.\end{aligned}$$

The expected sample size for the resampled θ_i's is therefore $M^{-1} n \int_{-\infty}^{\infty} f(x)\, dx$.

3.2 Random Variates via a Weighted Bootstrap

In cases where the bound M required in the preceding procedure is not readily available, we may still approximately resample from $h(\theta) = f(\theta)/\int f(\theta)\, d\theta$ as follows. Given θ_i, $i = 1, \ldots, n$, a sample from g, calculate $\omega_i = f(\theta_i)/g(\theta_i)$ and then

$$q_i = \omega_i \Big/ \sum_{j=1}^{n} \omega_j.$$

Draw θ^* from the discrete distribution over $\{\theta_1, \ldots, \theta_n\}$ placing mass q_i on θ_i. Then θ^* is approximately distributed according to h with the approximation "improving" as n increases. We provide a justification for this claim in a moment. However, first note that this procedure is a variant of the by now familiar bootstrap resampling procedure (Efron 1982). The usual bootstrap provides equally likely resampling of the θ_i, while here we have weighted resampling with weights determined by the ratio of f:g, again in agreement with intuition. See also Rubin (1988), who referred to this procedure as SIR (sampling/importance resampling).

Returning to our claim, suppose for convenience that θ is univariate. Under the customary bootstrap, θ^* has cdf

$$\begin{aligned}\Pr(\theta^* \leq a) &= \sum_{i=1}^{n} \frac{1}{n} 1_{(-\infty, a)}(\theta_i) \xrightarrow[n\to\infty]{} E_g 1_{(-\infty, a)}(\theta) \\ &= \int_{-\infty}^{a} g(\theta)\, d\theta\end{aligned}$$

so that θ^* is approximately distributed as an observation from $g(\theta)$. Similarly, under the weighted bootstrap, θ^* has cdf

$$\begin{aligned}\Pr(\theta^* \leq a) &= \sum_{i=1}^{n} q_i 1_{(-\infty, a)}(\theta_i) \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} \omega_i 1_{(-\infty, a)}(\theta_i)}{\frac{1}{n}\sum_{i=1}^{n} \omega_i} \xrightarrow[n\to\infty]{} \frac{E_g \frac{f(\theta)}{g(\theta)} \cdot 1_{(-\infty, a)}(\theta)}{E_g \frac{f(\theta)}{g(\theta)}} \\ &= \frac{\int_{-\infty}^{a} f(\theta)\, d\theta}{\int_{-\infty}^{\infty} f(\theta)\, d\theta} = \int_{-\infty}^{a} h(\theta)\, d\theta\end{aligned}$$

so that θ^* is approximately distributed as an observation from h. Note that the sample size under such resampling can be as large as desired. We mention one important caveat. The less h resembles g, the larger the sample size n will need to be in order that the distribution of θ^* well approximates h.

Finally, the fact that either resampling method allows h to be known only up to proportionality constant (i.e., only through f) is crucial, since in our Bayesian applications we wish to avoid the integration required to

standardize f. (Although, from the preceding, we can see that

$$n^{-1} \sum_{i=1}^{n} \omega_i$$

provides a consistent estimator of the normalizing constant

$$\int_{-\infty}^{\infty} f(\theta)\, d\theta$$

if such be required.)

4. BAYESIAN CALCULATIONS VIA SAMPLING–RESAMPLING

Both methods of the previous section may be used to resample the posterior (h) from the prior (g) and also to resample a second posterior (h) from a first (g). In this section we give details of both applications.

4.1 Prior to Posterior

How does Bayes's Theorem generate a posterior sample from a prior sample? For fixed x, define $f_x(\theta) = l(\theta; x)p(\theta)$. If $\hat{\theta}$ maximizes $l(\theta; x)$, let $M = l(\hat{\theta}; x)$. Then with $g(\theta) = p(\theta)$, we may immediately apply the rejection method of Section 3.1 to obtain samples from the density corresponding to the standardized f_x, which, from (1.1), is precisely the posterior density $p(\theta|x)$. Thus, we see that Bayes's Theorem, as a mechanism for generating a posterior sample from a prior sample, takes the following simple form: for each θ in the prior sample accept θ into the posterior sample with probability

$$\frac{f_x(\theta)}{Mp(\theta)} = \frac{l(\theta; x)}{l(\hat{\theta}; x)},$$

otherwise reject it.

The likelihood therefore acts as a resampling probability; those θ in the prior sample having high likelihood are more likely to be retained in the posterior sample. Of course, since $p(\theta|x) \propto l(\theta, x)p(\theta)$, we can also straightforwardly resample using the weighted bootstrap with

$$q_i = l(\theta_i; x) \Big/ \sum_{j=1}^{n} l(\theta_j; x).$$

Several obvious uses of this sampling–resampling perspective are immediate. Using large prior samples and iterating the resampling process for successive individual data elements—for two-dimensional θ, say—provides a simple pedagogic tool for illustrating the sequential Bayesian learning process, as well as the increasing concentration of the posterior as the amount of data increases. In addition, the approach provides natural links with elementary graphical displays (e.g., histograms, stem-and-leaf displays, boxplots to summarize univariate marginal posterior distributions, scatterplots to summarize bivariate posteriors). In general, the translation from functions to samples provides a wealth of opportunities for creative exploration of Bayesian ideas and calculations in the setting of computer graphical and exploratory data analysis (EDA) tools.

4.2 Posterior to Posterior

An important issue in Bayesian inference is sensitivity of inferences to model specification. In particular, we might ask:

1. How does the posterior change if we change the prior?
2. How does the posterior change if we change the likelihood?

In the density function/numerical integration setting, such sensitivity studies are rather off-putting, in that each change of a functional input typically requires one to carry out new calculations from scratch. This is not the case with the sampling–resampling approach, as we now illustrate in relation to the questions posed above.

In comparing two models in relation to the second question, we note that change in likelihood may arise in terms of:

1. change in distributional specification with θ retaining the same interpretation, for example, a location
2. change in data to a larger data set (prediction), a smaller data set (diagnostics), or a different data set (validation)

To unify notation, we shall in either case denote two likelihoods by $l_1(\theta)$ and $l_2(\theta)$. We denote two different priors to be compared in relation to the first question by $p_1(\theta)$ and $p_2(\theta)$. For complete generality, we shall consider changes to both l and p, although in any particular application we would not typically change both. Denoting the corresponding posterior densities by $\tilde{p}_1(\theta)$, $\tilde{p}_2(\theta)$,we easily see that

$$\tilde{p}_2(\theta) \propto \frac{l_2(\theta)p_2(\theta)}{l_1(\theta)p_1(\theta)} \cdot \tilde{p}_1(\theta). \tag{4.1}$$

Letting $v(\theta) = l_2(\theta)p_2(\theta)/l_1(\theta)p_1(\theta)$, we note that to implement the rejection method for (4.1) requires

$$\sup_{\theta} v(\theta).$$

In many examples this will simplify to an easy calculation. Alternatively, we may directly apply the weighted bootstrap method taking $g = \tilde{p}_1(\theta)$, $f = v(\theta)\tilde{p}_1(\theta)$, and $\omega_i = v(\theta_i)$. Resampled θ^* will then be approximately distributed according to f standardized, which is precisely $\tilde{p}_2(\theta)$.

Again, different aspects of the sensitivity of the posteriors to changes in inputs are easily studied by graphical examination of the posterior samples.

5. AN ILLUSTRATIVE EXAMPLE

To illustrate the passage, via Bayes's Theorem, from a prior sample to a posterior sample, we consider a two-

parameter problem first considered by McCullagh and Nelder (1989, sec. 9.3.3).

For $i = 1, 2, 3$, suppose that $X_{i1} \sim \text{Binomial}(n_{i1}, \theta_1)$ and $X_{i2} \sim \text{Binomial}(n_{i2}, \theta_2)$, conditionally independent given θ_1, θ_2, with n_{i1}, n_{i2} specified. Suppose further that the observed random variables are $Y_i = X_{i1} + X_{i2}$, $i = 1, 2, 3$, so that the likelihood for θ_1, θ_2 given $Y_1 = y_1$, $Y_2 = y_2$ and $Y_3 = y_3$ takes the form:

$$\prod_{i=1}^{3}\left[\sum_{j_i}\binom{n_{i1}}{j_i}\binom{n_{i2}}{y_i - j_i}\right.$$

$$\left.\times\ \theta_1^{j_i}(1-\theta_1)^{n_{i1}-y_i}\theta_2^{y_i-j_i}(1-\theta_2)^{n_{i2}-y_i+j_i}\right],$$

where $\max\{0, y_i - n_{i2}\} \le j_i \le \min\{n_{i1}, y_i\}$.

The data considered by McCullagh and Nelder are the following:

	i		
	1	2	3
n_{i1}	5	6	4
n_{i2}	5	4	6
y_i	7	5	6

For purposes of illustration, we take the joint prior distribution for θ_1, θ_2 to be uniform over the unit square. In accordance with the shift to a sampling perspective that constitutes our fundamental message in this article, Figure 1 presents a scatterplot of points uniformly drawn from the unit square, together with summary histograms confirming the uniform "shape" of the prior marginals for θ_1 and θ_2.

We now proceed to generate a posterior sample by resampling from the prior sample. For this illustration, the weighted bootstrap procedure was used and resulted in the posterior sample scatterplot shown in Figure 2, together with summary histograms of the posterior marginals for θ_1 and θ_2. General features of the posterior are easily identified from this picture—for example,

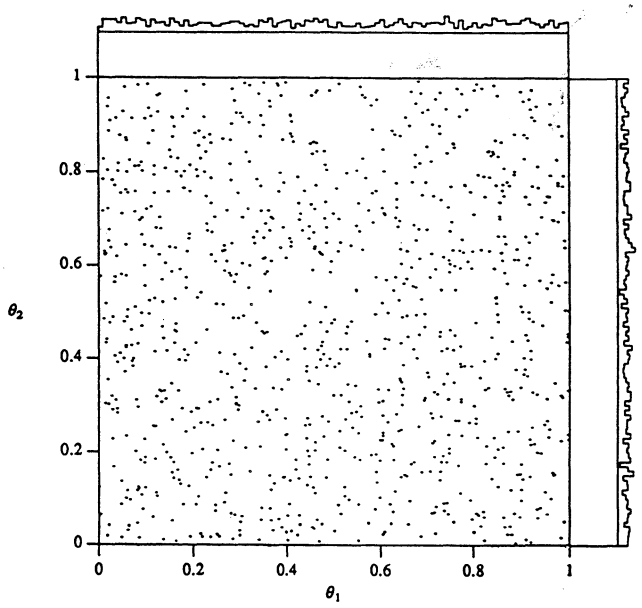

Figure 1. Sample from Prior.

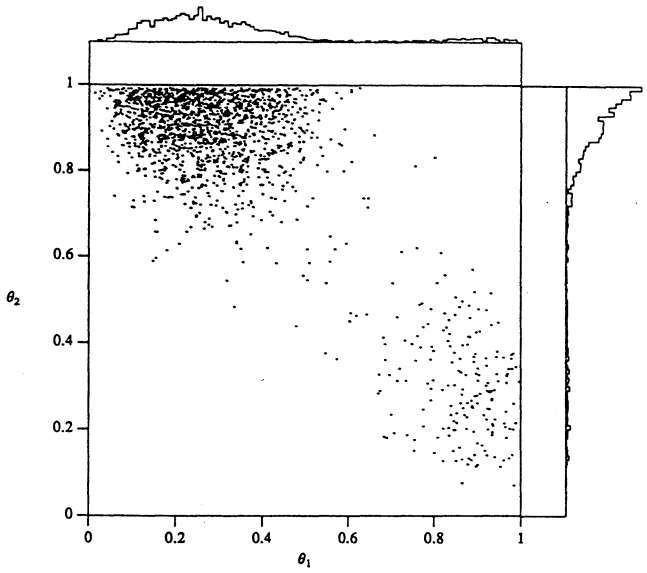

Figure 2. Sample from Posterior.

the marginal locations and spreads of inferences for the two parameters, together with the negative correlation and slight bimodality, which reflects the ambiguity resulting from observations in the form of sums of binomial outcomes.

Numerical summaries—in the form of posterior moments, quantiles, or whatever—are trivially obtained, if required, by forming corresponding sample quantities in the obvious way.

A further flexible and straightforwardly implemented feature of the sample-based approach is that posterior inferences can be trivially reexpressed in terms of any reparameterization of interest. For example, the logit (log-odds) reparameterization is often of interest in problems involving binomial data, so that, in the above, it might be of interest to recalculate the joint and mar-

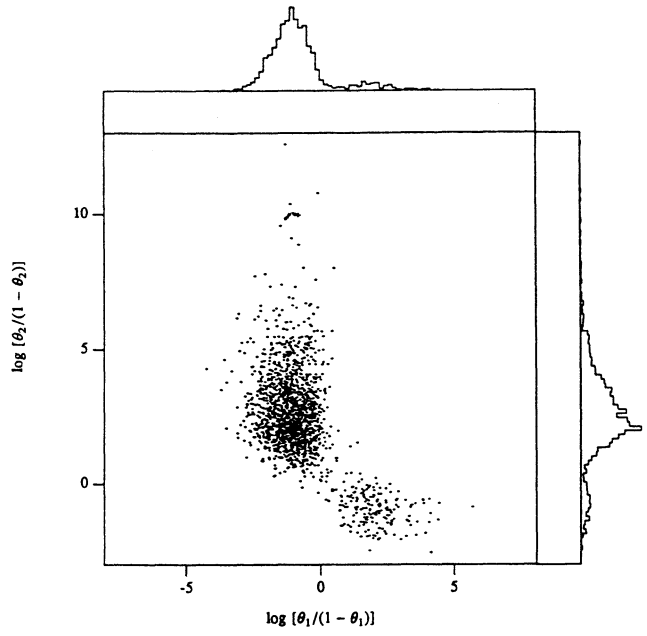

Figure 3. Sample from Transformed Posterior.

ginal posteriors for the parameters $\log [\theta_i/(1 - \theta_i)]$, $i = 1, 2$. The sample for θ_1, θ_2 translates directly into a sample from the logit transformed parameters and results in the summary picture given in Figure 3. We note that the forms of joint posterior revealed in Figures 2 and 3 are far from "nice" and would require subtle numerical handling by the nonsampling approaches cited in the Introduction.

So far as choice of sample sizes for initial and resamples is concerned, this will typically be a matter of experimentation with particular applications, having in mind the level of "precision" required from pictorial or numerical summaries. For example, in Figure 1 we display 1,000 sample points as an effective pictorial representation. However, the resampling ratio needs to be at least 1 in 10, with some 2,000 points plotted in Figures 2 and 3 to convey adequately these awkward posterior forms. The actual generated sample from the prior thus needed to be in excess of 20,000 points.

Clearly, there is considerable scope for more refined graphical outputs in terms of joint density contours, kernel density curves, and so on. We encourage readers to be creative in fusing EDA and graphical techniques with the sample-based approach to formal inference presented here.

[*Received May 1990. Revised January 1991.*]

REFERENCES

Efron, B. (1982), *The Bootstrap, Jackknife and Other Resampling Plans*, Philadelphia: Society of Industrial and Applied Mathematics.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Ripley, B. (1986), *Stochastic Simulation*, New York: John Wiley.

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

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., and Naylor, J. C. (1987), "Progress with Numerical and Graphical Methods for Bayesian Statistics," *The Statistician*, 36, 75–82.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C., and Dransfield, M. (1985), "The Implementation of the Bayesian Paradigm," *Communications in Statistics, Part A—Theory and Methods*, 14, 1079–1102.

Tierney, L., and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.