BAYES THEOREM FROM A SAMPLING-RESAMPLING PERSPECTIVE

BY

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

TECHNICAL REPORT NO. 445
JULY 31, 1991

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

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 theorem, relating the posterior distribution, $p(\theta|x)$, to the likelihood, $l(\theta;x)$, and the prior distribution, $p(\theta)$. 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—for example, 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; be it 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 shall address this problem by taking a new look at Bayes theorem from a sampling–resampling perspective. This will be seen to open the way both to easily implemented calculations and to essentially calculus-free insight into the mechanics and uses of Bayes 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 c.d.f. 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 which 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.

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 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, and the procedure is as follows (see, for example, Ripley, 1986, p.60):

(i) generate θ from $g(\theta)$;

(ii) generate u from uniform $(0, 1)$;

(iii) if $u \leq f(\theta)/Mg(\theta)$ accept θ; otherwise, repeat (i)–(iii).

Any accepted θ is then a random variate from $h(\theta) = f(\theta)/\int f(\theta)\, d\theta$.

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. Resulting sample size is random. Since it may be shown that the probability of acceptance of a random θ from g is M^{-1}, expected sample size for the resampled θ_i's is $M^{-1}n$.

3.2 Random variates via a weighted bootstrap

In cases where the bound M required in the above 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 $q_i = \omega_i / \sum_{j=1}^{n} \omega_j$. 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 to g, again in agreement with intuition.

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

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

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

$$P(\theta^* \leq a) = \sum_{i=1}^{n} q_i 1_{(-\infty,a]}(\theta_i) = \frac{\frac{1}{n}\sum_{i=1}^{n} \omega_i 1_{(-\infty,a]}(\theta_i)}{\frac{1}{n}\sum_{i=1}^{n} \omega_i}$$

$$\xrightarrow[n\to\infty]{} \frac{E_g \frac{f(\theta)}{g(\theta)} \cdot 1_{(-\infty,a]}(\theta)}{E_g \frac{f(\theta)}{g(\theta)}} = \frac{\int_{-\infty}^{a} f(\theta)\, d\theta}{\int_{-\infty}^{\infty} f(\theta)\, d\theta} = \int_{-\infty}^{a} h(\theta)\, d\theta$$

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.

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 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 f_x standardized, which, from (1.1), is precisely the posterior density $p(\theta|x)$. Thus, we see that Bayes 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 $q_i = l(\theta_i;x) / \sum_{j=1}^{n} l(\theta_j;x)$.

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, etc. 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 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:

how does the posterior change if we change the prior?
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

(i) change in distributional specification with θ retaining the same interpretation, e.g. a location,

(ii) 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

$$\tilde{p}_2(\theta) \propto \frac{l_2(\theta)p_2(\theta)}{l_1(\theta)p_1(\theta)} \cdot \tilde{p}_1(\theta). \tag{4.2}$$

Letting $v(\theta) = l_2(\theta)p_2(\theta)/l_1(\theta)p_1(\theta)$, we note that to implement the rejection method for (4.2) 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.

References

Efron B (1982). *The bootstrap, jackknife and other resampling plans*, SIAM, Philadelphia.

Ripley B (1986). *Stochastic simulation*, J Wiley & Sons, NY.

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, Theory and Methods* **14**, 1079–1102.

Smith A F M, Skene A M, Shaw J E H, Naylor J C (1987). Progress with numerical and graphical methods for Bayesian statistics, *The Statistician* **36**, 75–82.

Tierney L and Kadane J (1986). Accurate approximations for posterior moments and marginal densities, *J Amer Statist Assoc* **81**, 82–86.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 445	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Bayes Theorem From A Sampling-Resampling Perspective		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) A. F. M. Smith and A. E. Gelfand		8. CONTRACT OR GRANT NUMBER(s) N00014-89-J-1627
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE July 31, 1991 13. NUMBER OF PAGES 6
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bayesian inference; EDA; graphical methods; influence; posterior distribution; prediction; prior distribution; random variate generation; sampling-resampling techniques; sensitivity analysis; weighted-bootstrap.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Even to the initiated, statistical calculations based on Bayes 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.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED