note0402

Daniel Sorensen

July 6, 2024

Sensitivity Analysis

The output of a McMC study consists typically of samples $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$ from an approximate posterior distribution with density $p(\theta|y)$, where θ is the vector of parameters of the Bayesian model and y is the observed data vector. Expectations of functions $h(\theta)$ with respect to $p(\theta|y)$ can be estimated using the sample average:

$$\widehat{E}(h(\theta)) = \frac{1}{T}\sum_{i=1}^{T} h\left(\theta^{(i)}\right). \tag{1}$$

As part of the statistical analysis one may wish to study the robustness of the conclusions drawn from (1) to either different prior specifications or perturbations of the likelihood (such as a new functional form or part of the data could be omitted). Alternatively, interest may focus on modifying samples from one distribution to form samples from another distribution from which inferences can be drawn. This section describes two computational techniques that are useful to achieve these goals.

Importance Sampling

Imagine that (1) is an estimator of the expectation of $h(\theta)$ with respect to a posterior density

$$p_1(\theta|y) = c_1 p_1(\theta)\, p(y|\theta), \tag{2}$$

where c_1 is the typically unknown normalising constant, $p_1(\theta)$ is the prior density assigned to θ and $p(y|\theta)$ is the likelihood. Suppose that one wishes to study the consequences of changing the prior specification, such that the new posterior becomes

$$p_2(\theta|y) = c_2 p_2(\theta)\, p(\mathbf{y}|\theta). \tag{3}$$

Here, $p_2(\theta)$ is the density of the "new" prior distribution, and c_2 is the corresponding integration constant. Using the draws $\theta^{(i)}$ generated under the distribution with density $p_1(\theta|y)$, inferences about $h(\theta)$ under the new posterior with density $p_2(\theta|y)$ can be obtained without having to run the McMC procedure again, employing a computational technique

known as importance sampling (Hammersley and Handscomb, 1964). This is done by using $p_1(\theta|y)$ as importance sampling density. Thus, expectations under $p_2(\theta|y)$ can be obtained as follows

$$\begin{aligned} E_2[h(\theta)] &= \frac{\int h(\theta) \frac{p_2(\theta|y)}{p_1(\theta|y)} p_1(\theta|y)\, d\theta}{\int \frac{p_2(\theta|y)}{p_1(\theta|y)} p_1(\theta|y)\, d\theta} \\ &= \frac{\int h(\theta) \frac{c_2 p_2(\theta) p(\mathbf{y}|\boldsymbol{\theta})}{c_1 p_1(\theta) p(\mathbf{y}|\boldsymbol{\theta})} p_1(\theta|y)\, d\theta}{\int \frac{c_2 p_2(\theta) p(\mathbf{y}|\boldsymbol{\theta})}{c_1 p_1(\theta) p(\mathbf{y}|\boldsymbol{\theta})} p_1(\theta|y)\, d\theta} \\ &= \frac{\int h(\theta) w(\theta) p_1(\theta|y)\, d\theta}{\int w(\theta) p_1(\theta|y)\, d\theta}, \end{aligned} \tag{4}$$

where

$$w(\theta) = \frac{p_2(\theta)}{p_1(\theta)}.$$

Note that in the second line of (4) the ratio of constants of integration and the likelihood cancel out in the numerator and denominator. A consistent estimator of (4) based on (1) is

$$\widehat{E}_2[h(\theta)] = \frac{\sum_{i=1}^{T} h\left(\theta^{(i)}\right) w\left(\theta^{(i)}\right)}{\sum_{i=1}^{T} w\left(\theta^{(i)}\right)}, \tag{5}$$

where $\theta^{(i)}$, $(i = 1, 2, \ldots, T)$ are the draws from the distribution with density $p_1(\theta|y)$. The weight function w_i is equal to

$$w\left(\theta^{(i)}\right) = \frac{p_2\left[\theta^{(i)}\right]}{p_1\left[\theta^{(i)}\right]}.$$

If $w_i = 1$ for all i, $p_2(\theta|y) = p_1(\theta|y)$ and (4) is equal to (1).

The Monte Carlo standard error of (5) is

$$SE\left(\widehat{E}_2[h(\theta)]\right) = \frac{\sqrt{\sum_{i=1}^{T}\left[\left(h\left(\theta^{(i)}\right) - \widehat{E}_2[h(\theta)]\right) w\left(\theta^{(i)}\right)\right]^2}}{\sum_{i=1}^{T} w\left(\theta^{(i)}\right)} \tag{6}$$

(Geweke, 1989). The presence of large relative weights leads to poor estimates.

Moments and quantiles under the new posterior distribution can be obtained along the same lines, using the draws from the original posterior distribution. For instance

$$\begin{aligned} \widehat{Var}_2[h(\theta)] &= \widehat{E}_2\left[h^2(\theta)\right] - \left[\widehat{E}_2[h(\theta)]\right]^2 \\ &= \frac{\sum_{i=1}^{T} h^2\left[\theta^{(i)}\right] w\left(\theta^{(i)}\right)}{\sum_{i=1}^{T} w\left(\theta^{(i)}\right)} - \left[\widehat{E}_2[h(\theta)]\right]^2 \end{aligned} \tag{7}$$

and

$$\widehat{\Pr}_2\left[h\left(\theta\right) < c\right] = \frac{\sum_{i=1}^{T} I\left[h\left(\theta^{(i)}\right) < c\right] w\left(\theta^{(i)}\right)}{\sum_{i=1}^{T} w\left(\theta^{(i)}\right)}, \tag{8}$$

where subscript 2 indicates that inferences are being drawn from the posterior distribution with density $p_2\left[\theta|y\right]$ and $\theta^{(i)}$ is the ith sample drawn from the posterior distribution with density $p_1\left[\theta|y\right]$.

Often, it can be computationally advantageous to fit a particular model elicited under a certain prior or likelihood specification. However, the analyst may have in mind an alternative model which is less tractable computationally. The approach described above provides a powerful tool for doing this in a rather straightforward manner. This is illustrated in the following example, taken from Sorensen and Gianola (2002).

Example: Inferences From Two Beta Distributions

Suppose n independent draws are made from a Bernoulli distribution with unknown probability of success θ. Let x denote the number of successes and y the number of failures. The likelihood is

$$p\left(x|\theta, n\right) \propto \theta^x \left(1-\theta\right)^y. \tag{9}$$

The experimenter wishes to perform the Bayesian analysis under two different sets of prior assumptions. The first model assumes a uniform prior distribution for θ, $Un\left(0,1\right)$:

$$p_1\left(\theta\right) = 1, \quad 0 \leq \theta \leq 1. \tag{10}$$

Under this prior, the posterior density is proportional to (9)

$$p_1\left(\theta|x, n\right) \propto \theta^x \left(1-\theta\right)^y, \tag{11}$$

which is recognized as the density of a beta distributed random variable with parameters $x+1, y+1$, that is $Be\left(\theta|x+1, y+1\right)$. The second model assumes the same likelihood, but the prior distribution for θ is beta, with parameters a and b. The posterior density is now

$$p_2\left(\theta|x, n\right) \propto \theta^{a+x-1} \left(1-\theta\right)^{b+y-1}, \tag{12}$$

which is the density $Be\left(\theta|a+x, b+y\right)$. In this example, the form of the posterior distribution is known under either prior, so it is straightforward to draw inferences from (11) or from (12). To illustrate, independent samples will be drawn from (11), and then importance sampling will be used to obtain inferences based on (12), using the draws from (11). Further, the Monte Carlo-based estimates will then be compared with exact results. The results for $\theta = 0.8$, obtained with $n = x + y = 5$ or 15, are shown in Table 1, for three importance sampling sample sizes. The focus of inference is on the posterior mean and variance, and on the probability that the value of θ lies between 0.75 and 0.85. In

			Mean × 10		Variance × 10^2		Probability	
S	x	y	Exact	IS	Exact	IS	Exact	IS
4	4	1	6.364	6.356	1.9284	1.9757	0.1742	0.1710
100	4	1	6.364	6.361	1.9284	1.9322	0.1742	0.1728
1000	4	1	6.364	6.365	1.9284	1.9286	0.1742	0.1743
4	12	3	7.143	7.134	0.9276	0.9515	0.3155	0.3004
100	12	3	7.143	7.141	0.9276	0.9283	0.3155	0.3124
1000	12	3	7.143	7.143	0.9276	0.9277	0.3155	0.3157

Table 1: Comparison betwen exact results and estimates based on importance sampling (IS). S: number of samples in thousands; x: number of successes; y: number of failures; Probability: posterior probability that the binomial parameter takes a value between 0.75 and 0.85. From Sorensen and Gianola (2002).

the model that provides the basis of inference, $p_2(\theta|x,n)$, the parameters of the Beta prior are $a = b = 3$. The results in the table illustrate that the estimator is consistent: as the number of samples increases from 4000 to 1 million, the estimates based on (5), (7), and (8) converge to the true values.

When the probability to be estimated is small, a larger number of importance samples must be drawn to achieve the same level of precision. For example, the true probability that θ lies between 0.3 and 0.4, based on $p_2(\theta|x,n)$, is 15.68×10^{-4}. Estimates obtained with sample sizes of four thousand, one hundred thousand and one million were 12.10×10^{-4}, 14.66×10^{-4}, and 15.75×10^{-4}, respectively.

While in this example the importance sampling approach performs satisfactorily, in higher-dimensional problems the relative weights

$$\frac{w\left(\theta^{(i)}\right)}{\sum_{i=1}^{n} w\left(\theta^{(i)}\right)}$$

may be concentrated on a small number of samples. As a consequence, the Monte Carlo sampling error associated with estimates of posterior features is likely to be large. A larger effective sample size is required in order to mitigate this drawback.

Sampling Importance Resampling

Rubin (1987), Rubin (1988), Smith and Gelfand (1992) and Albert (2009) describe how another computational technique known as importance resampling (or sampling importance resampling, SIR) can be used to modify samples from one distribution to form samples from another distribution. Suppose that samples $\theta_1 = \left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}\right)$ are available from a distribution with density p_1 and there is interest in transforming θ_1 into samples from a distribution with density p_2. This is achieved as follows:

1. compute weights $w^{(j)} = \frac{p_2\left(\theta^{(j)}\right)}{p_1\left(\theta^{(j)}\right)}, \quad j = 1, 2, \ldots, T,$

2. convert the weights to probabilities $w_j = \frac{w^{(j)}}{\sum_{j=1}^{T} w^{(j)}},$

3. sample with replacement from the vector θ_1 with probabilities w_j. This results in a new vector θ_2 that is an approximate sample from the distribution with density p_2.

This procedure, named sampling importance resampling (SIR), is a weighted bootstrap where the elements of θ_2 are sampled from the sample θ_1 with unequal probabilities.

`note1001.pdf` shows an application where samples from a posterior distribution $[\theta|y]$ are transformed to samples from a posterior distribution $[\theta|y_{-i}]$, where θ is the vector of parameters of the Bayesian model and data vector $y = (y_i, y_{-i})$, $i = 1, 2, \ldots, N$, with y_{-i} equal to y with the ith observation excluded. Following the three steps above, using $p_2(\theta) = p(\theta|y_{-i})$, $p_1(\theta) = p(\theta|y)$, results in weights of the form

$$w_{ij} = \frac{p^{-1}\left(y_i|\theta^{(j)}\right)}{\sum_{j=1}^{T} p^{-1}\left(y_i|\theta^{(j)}\right)}, \; i = 1, 2, \ldots, N; \; j = 1, 2, \ldots, T. \tag{13}$$

Sampling with replacement with probabilities $(w_{i1}, w_{i2}, \ldots, w_{iT})$ from the vector $\theta_1 = \left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}\right)$ drawn from $[\theta|y]$, generates an approximate sample from the posterior distribution $[\theta|y_{-i}]$. Indeed, for example when θ is a scalar on the real line,

$$\begin{aligned} \Pr(\theta \le a|y_{-i}) &= \frac{\int_{-\infty}^{a} p(\theta|y_{-i})\, d\theta}{\int_{-\infty}^{\infty} p(\theta|y_{-i})\, d\theta} = \frac{\int_{-\infty}^{a} \frac{p(\theta|y_{-i})}{p(\theta|y)} p(\theta|y)\, d\theta}{\int_{-\infty}^{\infty} \frac{p(\theta|y_{-i})}{p(\theta|y)} p(\theta|y)\, d\theta} \\ &= \frac{\frac{p(y)}{p(y_{-i})}}{\frac{p(y)}{p(y_{-i})}} \frac{\int_{-\infty}^{a} \frac{1}{p(y_i|\theta)} p(\theta|y)\, d\theta}{\int_{-\infty}^{\infty} \frac{1}{p(y_i|\theta)} p(\theta|y)\, d\theta} = \frac{\int_{-\infty}^{a} \frac{1}{p(y_i|\theta)} p(\theta|y)\, d\theta}{\int_{-\infty}^{\infty} \frac{1}{p(y_i|\theta)} p(\theta|y)\, d\theta}, \end{aligned} \tag{14}$$

so the distribution $[\theta|y_{-i}]$ is generated by appropriate weighting of the distribution $[\theta|y]$, as indicated in the bottom right hand side of (14). A Monte Carlo consistent estimator of (14) is

$$\begin{aligned} \widehat{\Pr}(\theta \le a|y_{-i}) &= \frac{\sum_{j=1}^{T} \frac{1}{p\left(y_i|\theta^{(j)}\right)} I\left(\theta^{(j)} \le a\right)}{\sum_{j=1}^{T} \frac{1}{p\left(y_i|\theta^{(j)}\right)}} \\ &= \sum_{j=1}^{T} w_{ij} I\left(\theta^{(j)} \le a\right), \quad i = 1, 2, \ldots, N, \end{aligned} \tag{15}$$

where w_{ij} is given by (13), $\theta^{(j)}$ is the jth draw from $[\theta|y]$ once the McMC has converged to its stationary distribution and $I(\cdot)$ is the indicator function that takes the value 1 if the argument is satisfied and 0 otherwise.

`note1001.pdf` develops these ideas to obtain Bayesian estimators of leave-one-out cross-validation running through the data only once and an example is provided. Attention is drawn to the fact that the estimates using the weighted bootstrap are sensitive to the presence of large values of the weights and some form of smoothing can produce less noisy estimates.

References

Albert, J. (2009). *Bayesian Computation with R*. Springer.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Hammersley, J. M. and D. C. Handscomb (1964). *Monte Carlo Methods*. Wiley.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. Discussion of Tanner and Wong. *Journal of the American Statistical Association 82*, 543–546.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, pp. 395–402. Oxford University Press.

Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician 46*, 84–88.

Sorensen, D. and D. Gianola (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag. 740 pp., Reprinted with corrections, 2006.