Stat314/461Term 4: Simple simulation methods for non-standard densities: Importance sampling

Patrick Graham

September, 2021

Reminder: Basic problem of Bayesian computation

$$p(\theta|\mathbf{Y}^{obs}) = \frac{p(\mathbf{Y}^{obs}|\theta)p(\theta)}{\int p(\mathbf{Y}^{obs}|\theta)p(\theta)\,\mathrm{d}\theta} \tag{1}$$

- We specify $p(\theta)$
- We derive $p(\mathbf{Y}^{obs}|\theta)$ from our data model
- Hence we can always write an expression for the numerator of (1).
- But to do anything useful with (1) we need to be able to integrate the numerator and this may be difficult to do explicitly in realistic multi-parameter, non-conjugate problems.
- Modern Bayesian computation makes extensive use of Monte Carlo methods to *simulate* the posterior distribution.

Importance sampling

- Suppose $p(\theta|\mathbf{Y}^{obs}) \propto p(\mathbf{Y}^{obs}|\theta)p(\theta)$ does not correspond to a standard distribution, we cannot simulate directly from it and cannot easily integrate it.
- We can compute the unnormalised posterior $q(\theta|\mathrm{Y}^{obs}) = p(\mathbf{Y}^{obs}|\theta)p(\theta)$.
- We can easily simulate from an approximation $g(\theta)$
- Suppose we would like to compute $E(h(\theta)|\mathbf{Y}^{obs})$ for some function $h(\theta)$, e.g. $h(\theta) = \theta$ or $h(\theta) = (\theta - E(\theta|\mathbf{Y}^{obs}))^2$

$$E(h(\theta)|\mathbf{Y}^{obs}) = \int h(\theta)p(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta \tag{2}$$

$$= \frac{\int h(\theta)q(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta}{\int q(\theta|\mathbf{Y}^{obs})\,\mathrm{d}\theta} \tag{3}$$

$$= \frac{\int \left[h(\theta)q(\theta|\mathbf{Y}^{obs})/g(\theta)\right] g(\theta)\,\mathrm{d}\theta}{\int \left[q(\theta|\mathbf{Y}^{obs})/g(\theta)\right] g(\theta)\,\mathrm{d}\theta} \tag{4}$$

Importance sampling algorithm when only the unnormalised posterior is known

$$E(h(\theta)|\mathbf{Y}^{obs}) = \frac{\int \left[h(\theta)q(\theta|\mathbf{Y}^{obs})/g(\theta)\right] g(\theta)\,\mathrm{d}\theta}{\int \left[q(\theta|\mathbf{Y}^{obs})/g(\theta)\right] g(\theta)\,\mathrm{d}\theta} \tag{5}$$

We can apply a form of Monte Carlo integration to evaluate the numerator and denominator of (5) Suppose we want $E(h(\theta)|\mathbf{Y}^{obs})$

for (i in $1 \ldots \mathrm{M}$)

1. draw $\theta_{(i)}$ from $g(\theta)$
2. compute $r(\theta_{(i)}) = q(\theta_{(i)}|\mathbf{Y}^{obs})/g(\theta_{(i)})$
3. compute $h(\theta_{(i)})$
4. store $r(\theta_{(i)}), h(\theta_{(i)})$

$$E(h(\theta|\mathbf{Y}) \approx \frac{\sum_i h(\theta_{(i)})r(\theta_{(i)})}{\sum_i r(\theta_{(i)})} \tag{6}$$

Effective Monte Carlo sample size for importance sampling

$r(\theta_{(i)}) = q(\theta_i|\mathbf{Y}^{obs})$.
Let $\tilde{r}(\theta_i) = r(\theta_i)/\sum_{i=1}^{M} r(\theta_i))$ denote the normalised importance ratios $(\sum_{i=1}^{M})\tilde{r}(\theta_i) = 1)$ then the effective Monte Carlo sample size is

$$N_{\text{eff}} = \frac{1}{\sum_{i=1}^{M} (\tilde{r}(\theta_i))^2}. \tag{7}$$

$N_{\text{eff}} \leq M$; equality holds if the importance weights are constant
$N_{\text{eff}} \ll M$; if weights are highly variable, e.g a few very large weights.

Approximating the Monte Carlo error for importance sampling

Since we have an approximation to the effective Monte Carlo sample size we can also obtain an approximation to the Monte Carlo error.

- $E(\theta|\mathbf{Y}^{obs}) \approx \hat{E}(\theta|\mathbf{Y}^{obs}) = \sum_i \tilde{r}(\theta_i)\theta_i$
- $V(\theta|\mathbf{Y}^{obs}) \approx \hat{V}(\theta|\mathbf{Y}^{obs}) = \sum_i \tilde{r}(\theta_i)(\theta_i - \hat{E}(\theta|\mathbf{Y}^{obs}))^2$
- $sd(\theta|\mathbf{Y}^{obs}) \approx \hat{sd}(\theta|\mathbf{Y}^{obs}) = \sqrt{\hat{V}(\theta|\mathbf{Y}^{obs})}$
- $\mathrm{MCerror} \approx \hat{sd}(\theta|\mathbf{Y}^{obs})/\sqrt{N_{\mathrm{eff}}}$

And we recall that the $\tilde{r}(\theta_i$ are the normalised importance ratios (or weights and, so, sum to one).

Importance sampling when the approximating density is the prior

Note if $g(\theta)$ is the prior $p(\theta)$ then

$$
\begin{aligned}
r(\theta) &= \frac{p(\theta|\mathbf{Y}^{obs})}{p(\theta)} \\
&= \frac{p(\mathbf{Y}^{obs}|\theta)p(\theta)}{p(\theta)} \qquad (8) \\
&= p(\mathbf{Y}^{obs}|\theta) = \qquad (9)
\end{aligned}
$$

i.e the likelihood. The prior weighted by the likelihood is the posterior!

Comments on importance sampling algorithm

- The formulation is very general. If we are interested particular posterior probabilities, e.g. $\Pr(a \leq \theta \leq b | \mathbf{Y}^{obs})$ just define $h(\theta) = I(a \leq \theta \leq b)$.
- For most practical purposes, can just treat the *weighted* sample of θ's as a sample from the posterior. Need to keep in mind the importance sample just provides an *approximation* to the posterior
- Plotting posterior densities or histograms is a bit awkward because of the weights. - A simple solution is to resample the original θ sample with probability proportional to the importance sampling ratio. Then plot the resulting sample.

More comments on importance sampling

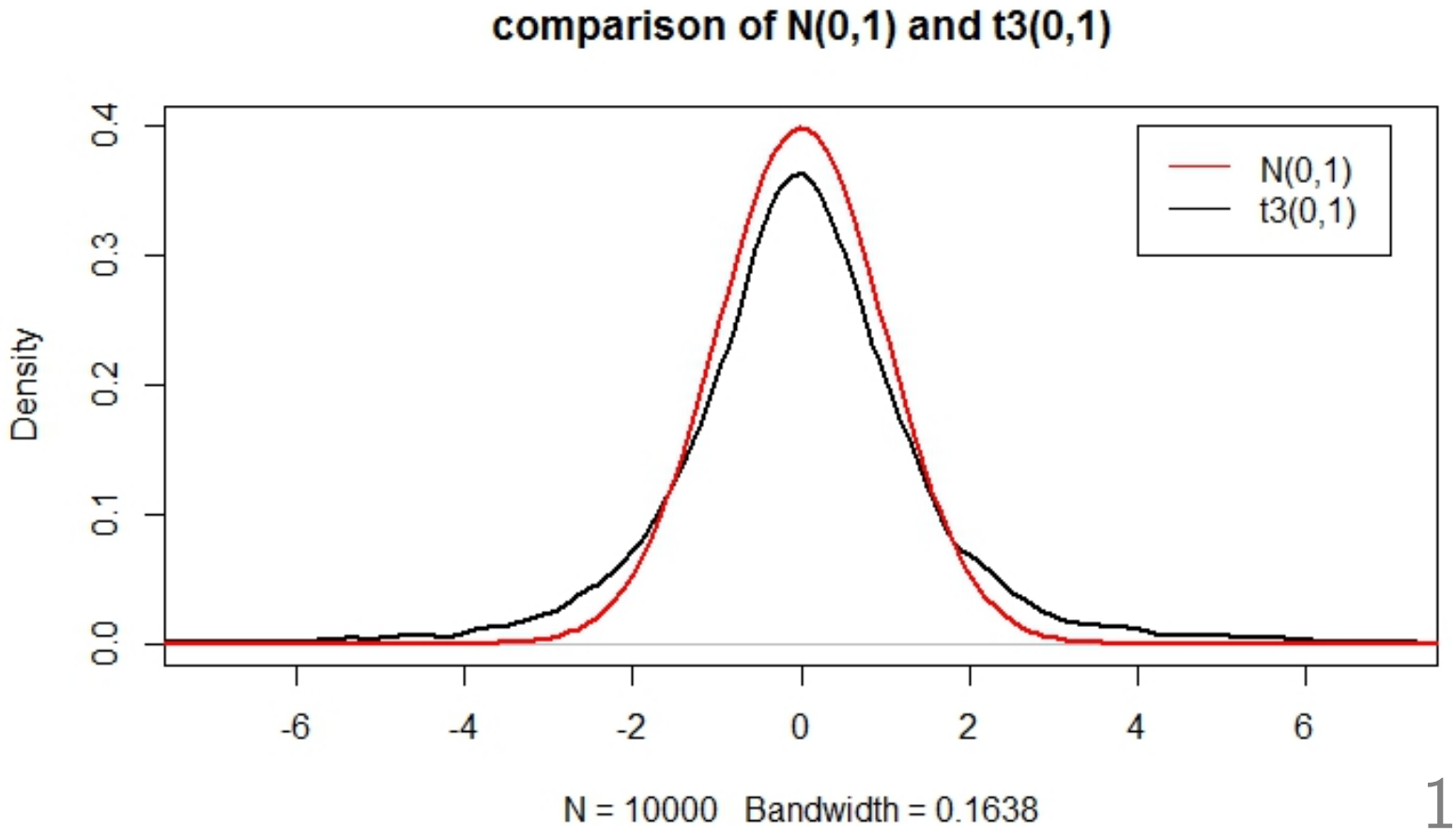
- If the distribution of importance sampling weights is very uneven with a small number of θ values having very large weights, then most of the information about the posterior will be concentrated on only a few sample points. Not ideal and means the effective Monte Carlo sample size will be much less than the nominal size.
- It is important and helpful to plot a histogram of log importance weights before proceeding to inference. Concentrate on the distribution of largest importance sampling weights, e.g top 30%.

History and applications of importance sampling

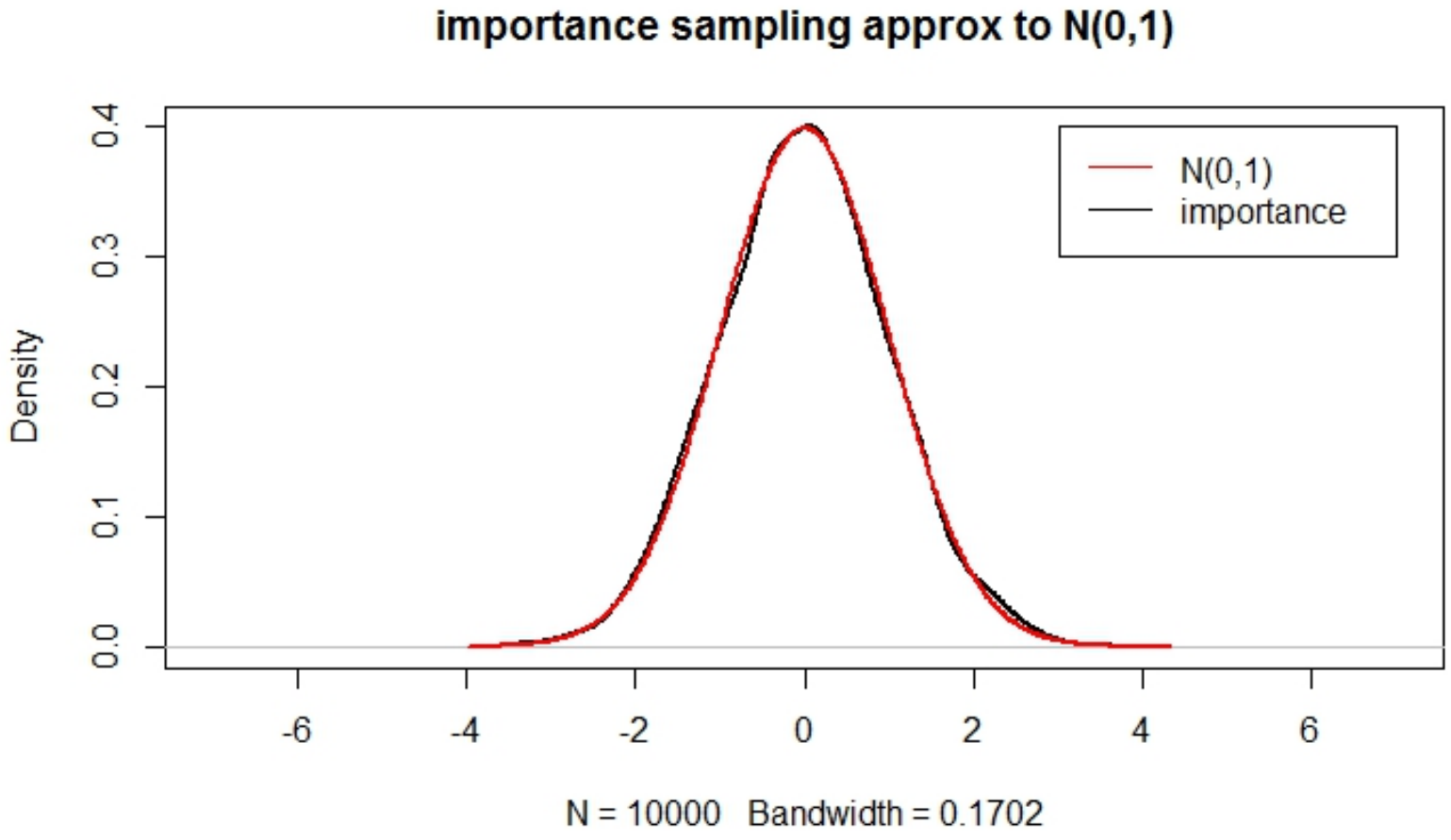
- Prior to the MCMC revolution beginning around 1990, importance sampling was an active area of research and practice in Bayesian statistics, with various clever ways of forming approximations to the posterior developed.
- It still features today as a reasonable approach for simple problems and as a component of more advanced methods such as Sequential Monte Carlo.
- Recently, a form of importance of sampling (Pareto smoothed importance sampling) has found application in the development of "leave one out cross-validation" for model comparison and selection. Here the posterior needs to repeatedly re-computed on data with one observation dropped each time. Importance sampling provides an efficient means of doing that.
  `https://arxiv.org/pdf/1507.02646.pdf`

Examples of good and bad importance samplers

- using a t with small df to approximate a Normal distribution is good.
- using a Normal to approximate a t with low degrees of freedom is not so good.



Importance sampling approximation to normal based on t3(0,1) approximation



Application of importance sampling to the "unknown N, known p" problem

see separate importance sampling code `importance_sampling_examples.2021.Rmd`