Lecture 7: Intro to Bayesian Computation

James D. Wilson

MSAN 628

Computational Statistics

Outline

- Monte Carlo Simulations
- Numeric Integration
- Rejection Sampling
- Importance Sampling

Motivation

- Bayesian computation revolves around two primary calculations:
  - Posterior distributions: $p(\theta \mid y)$
  - Posterior predictive distributions $p(\tilde{y} \mid y)$
- So far, we have dealt with posteriors where values can be analytically calculated (e.g., conjugate families)
- In many cases, however, distributions are much more difficult and cannot be written down or are challenging to simulate from. In such cases, we can use Monte Carlo simulations to approximate values from a calculated density.

The Setting

Observe data $\mathbf{y}$ and we propose a (possibly multivariate) prior distribution $\pi(\theta)$ as well as a data generating density $f(\mathbf{y} \mid \theta)$.

- Target distribution: the distribution to be simulated from - the posterior distribution $p(\theta \mid \mathbf{y})$
- Unnormalized density: an easily computable function $q(\theta \mid \mathbf{y})$ for which $q(\theta \mid \mathbf{y})/p(\theta \mid \mathbf{y})$ is a function of only $\mathbf{y}$.
  - Typically, we use the kernel of the density (which we've been using all year so far):

$$q(\theta \mid \mathbf{y}) = \pi(\theta) f(\mathbf{y} \mid \theta)$$

Note on Computation

For any computation of densities, we will typically **compute logarithms** rather than the density itself.

- This helps avoid overflow and underflow of computational storage
- Exponentiation should only be performed at the *last* step of computation!
- **Example**: Rather than computing the ratio of two densities $q(\theta \mid \mathbf{y})/p(\theta \mid \mathbf{y})$, we can simply calculate the log of the ratio:

$$\log(q(\theta \mid \mathbf{y})/p(\theta \mid \mathbf{y})) = \log(q(\theta \mid \mathbf{y})) - \log(p(\theta \mid \mathbf{y}))$$

  Then, if you need the ratio of the two, exponentiate after this computation has been done!

Numerical Integration

- We begin with the task of integrating over values of the posterior distribution given that **we know** its distributional form.
- Important quantities to consider:
  - **Expectations**: the posterior expectation of any function $h(\theta)$:

    $$\mathbb{E}[h(\theta) \mid \mathbf{y}] = \int h(\theta)p(\theta \mid \mathbf{y})d\theta$$

    which is an integral with the same number of dimensions as θ.
  - **Posterior Predictive Distributions**: for new data $\tilde{y}$, we want

    $$p(\tilde{y} \mid \mathbf{y}) = \int f(\tilde{y} \mid \theta)p(\theta \mid \mathbf{y})d\theta$$

Simulation Method for Numerical Integration

- First simulate S samples $\theta^1, \ldots, \theta^S$ from the posterior distribution $p(\theta \mid \mathbf{y})$
- **Approximate Expectation**:

$$\mathbb{E}[h(\theta) \mid \mathbf{y}] = \int h(\theta)p(\theta \mid \mathbf{y})d\theta \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^s)$$

- **Approximate Predictions**:

$$p(\widetilde{y} \mid \mathbf{y}) = \int f(\widetilde{y} \mid \theta)p(\theta \mid \mathbf{y})d\theta \approx \frac{1}{S}\sum_{s=1}^{S} f(\widetilde{y} \mid \theta^s)$$

Example: Normal-Normal model in action

- Suppose that $\theta \sim N(0, 1)$ and $y \mid \theta \sim N(\theta, 2)$
- For n observations of y, we know (thanks wikipedia) that $\theta \mid \mathbf{y} \sim N(\mu, \sigma^2)$, where

$$\mu = \frac{\sum_{i=1}^{n} y_i/2}{1 + n/2}$$

$$\sigma^2 = (1 + n/2)^{-1}$$

Goals:

- Approximate $\mathbb{E}[\log(\theta) \mid \mathbf{y}]$
- Simulate the posterior predictive distribution for $\widetilde{y} \mid \mathbf{y}$

Go to $\mathbb{R}$ code on Canvas here:

Notes about Numerical Integration

- The estimate depends on the randomness of random number generators
- The estimate improves as one draws more samples. That is, as $S \to \infty$.
- This idea works due to the law of large numbers in probability!

Deterministic Methods for Numerical Integration

- In some cases, we can come up with an "intelligent" way to weight each simulated value.
- That is, suppose that we devise a weight w_s for each sample $s = 1, \ldots, S$. Then we can estimate expectations as:

$$\mathbb{E}[h(\theta) \mid \mathbf{y}] \approx \frac{1}{S} \sum_{s=1}^{S} w_s h(\theta^s) p(\theta^s \mid \mathbf{y})$$

- In general, this has lower variance than other simulation-based methods.
- However, this relies on smart ways of choosing weights. Some already exist, including: quadrature rules like Simpson's rule, etc.

Rejection Sampling

Goal: Simulate from $p(\theta \mid \mathbf{y})$

Key Ingredient: need a positive proposal function $g(\theta)$ defined for all θ such that $p(\theta \mid \mathbf{y}) > 0$ that satisfies the following:

- $\int g(\theta) d\theta = C < \infty$. (must have a finite integral)
- There exists a finite bound $M < \infty$ such that for all θ,

$$\frac{p(\theta \mid \mathbf{y})}{g(\theta)} \leq M$$

Note: The value $\frac{p(\theta|\mathbf{y})}{g(\theta)}$ is known as the importance ratio.

Rejection Sampling

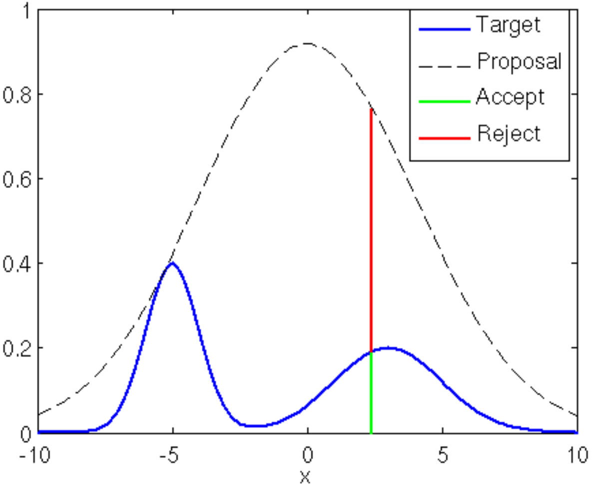
Algorithm

Goal: Simulate from $p(\theta \mid \mathbf{y})$

1. Sample θ at random from the density proportional to $g(\theta)$. In other words, simulate from $g(\theta)/C$
2. Accept the sample with probability $\dfrac{p(\theta \mid \mathbf{y})}{Mg(\theta)}$. If the draw is rejected, repeat step 1.

Key Takeaway: this requires a good approximation of $p(\theta \mid \mathbf{y})$ which has a closed form density.

Rejection Sampling Illustration

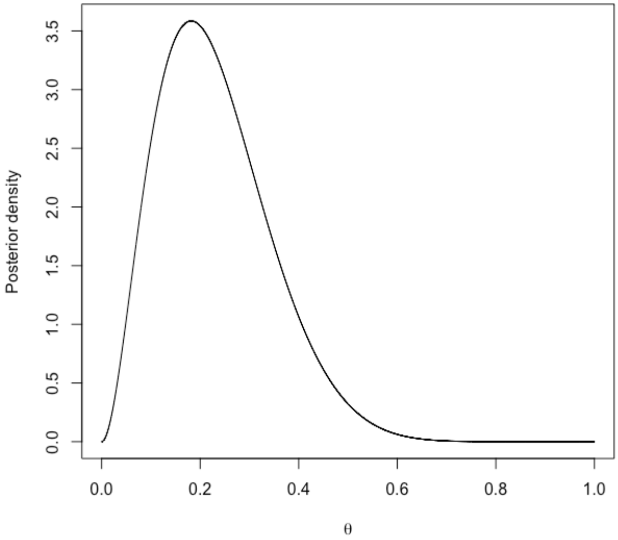


Notes about Rejection Sampling

- A good approximate density $g(\theta)$ for rejection sampling should be roughly proportional to $p(\theta \mid \mathbf{y})$.
- The ideal situation is $g \propto p$, in which case with a suitable value of M, we accept every draw with probability 1.
- Benefit of rejection sampling: self-monitoring - if the method is not working efficiently, very few simulated values will be accepted!
- Once again, the idea here is that we have a functional form of $p(\theta \mid \mathbf{y})$ but want to simulate from the density.

Example of Rejection Sampling

Suppose that we want to simulate from the below density:



Example of Rejection Sampling

- We have to pick a known (and integrable) $g(\theta)$ so that $Mg(\theta) \geq p(\theta \mid \mathbf{y})$

  Natural choice:

  - Set $M = \max_\theta(p(\theta \mid \mathbf{y}))$
  - Let $g(\theta) = 1$ (Uniform distribution on [0,1]).

- This choice satisfies our needed condition.

See the `R` script for more details.

Importance Sampling

- **Setting**: Calculate expectations from posterior densities that *you do not know*!
- In particular, we would like to calculate $\mathbb{E}[h(\theta) \mid \mathbf{y}]$, but we do not know $p(\theta \mid \mathbf{y})$.
- How is this even possible?!

Importance Sampling

Idea: Use a density of θ that we *do* know: $g(\theta)$ and the unnormalized density $q(\theta \mid \mathbf{y})$:

Fact:

$$\mathbb{E}[h(\theta) \mid \mathbf{y}] = \frac{\int h(\theta)q(\theta \mid \mathbf{y})d\theta}{\int q(\theta \mid \mathbf{y})d\theta}$$
$$= \frac{\int [h(\theta)q(\theta \mid \mathbf{y})/g(\theta)]g(\theta)d\theta}{\int [q(\theta \mid \mathbf{y})/g(\theta)]g(\theta)d\theta}$$

Importance Sampling

The previous quantity can be estimated with Monte Carlo methods using S draws $\theta^1, \ldots, \theta^S$ from $g(\theta)$ by the expression:

$$\frac{\sum_{s=1}^{S} h(\theta^s) w(\theta^s)}{\sum_{s=1}^{S} w(\theta^s)},$$

where

$$w(\theta^s) = \frac{q(\theta^s \mid \mathbf{y})}{g(\theta^s)}$$

are the importance ratios or importance weights.

Notes about Importance Sampling

- Relies on choice of proposal distribution $g(\theta)$
- If $g(\theta)$ can be chosen such that hq/g is roughly constant, then fairly precise estimates of the integral can be obtained.
- Worst case scenario: importance ratios are small with high probability but are large with low probability (when hq has wide tails compared to g).

Effective Sample Size for Importance Sampling

- A rough estimate of the number of samples needed for convergence is given by the effective sample size.
- When the variance of the weights is finite, an estimate of the effective sample size is given by

$$S_{eff} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2},$$

where

$$\tilde{w}(\theta^s) = S * w(\theta^s) / \sum_{s=1}^{S} w(\theta^s)$$

- Note that this estimate is itself noisy, so it acts only as a rough guide!

Example of Importance Sampling

- Suppose that our (unknown) posterior distribution is $\theta \mid \mathbf{y} \sim N(1, 1)$
- We get some idea of our data in that it looks a kind of normal but with "fatter" tails
- So, we propose a distribution $g(\theta) \sim t_3$, the t-distribution with 3 degrees of freedom and non-centrality parameter $\mu = \sqrt{4/3}$ (to match the mode of the posterior distribution)
- **Goal**: Approximate $E[\theta \mid \mathbf{y}]$ and $\text{Var}(\theta \mid \mathbf{y})$.

See the `R` script.

Next Time

- Markov Chain Basics
- Introduction to Markov Chain Simulation
- Gibbs Sampler
- Metropolis Hastings and Metropolis Algorithms