Statistics 568 Bayesian Analysis

Intro to Bayesian Computation

Ruobin Gong

Department of Statistics
Rutgers University

Week 6 - 02/25/21

Targets of Bayesian computation

Bayesian computation revolves around the computation of two targets:

1. posterior distribution $p(\theta \mid y)$, and
2. posterior predictive distribution $p(\tilde{y} \mid y)$.

Notation

- The **target distribution** is denoted as $p(\cdot)$
- The **unnormalized target distribution** is denoted by $q(\cdot)$
  - $\int q(a)da < \infty$;
  - $q(\cdot) \propto p(\cdot)$.
- The **proposal distribution** is denoted by $g(\cdot)$.

Note. When discussing computation for a single model, the posterior's dependence on y is not important, so we sometimes use the generic $p(\cdot)$ and $q(\cdot)$ instead of $p(\cdot \mid y)$ and $q(\cdot \mid y)$ whenever no confusion arises.

Approximating expectations

Let $f(\theta)$ be a quantity of interest. Its posterior expectation

$$E\left(f(\theta) \mid y\right) = \int f(\theta) p(\theta|y) d\theta,$$

$$\text{where} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

We can* easily evaluate $p(y|\theta)p(\theta)$ for any θ, but the integral $\int p(y|\theta)p(\theta)d\theta$ is usually difficult.

*. Approximate Bayesian Computation.

Approximating expectations

We can use the unnormalized posterior $q(\theta|y) = p(y|\theta)p(\theta)$, for example, in

- Grid (equal spacing) evaluation with self-normalization

$$E\left(f(\theta) \mid y\right) \approx \frac{\sum_{s=1}^{S}\left[f(\theta^{(s)})q(\theta^{(s)}|y)\right]}{\sum_{s=1}^{S} q(\theta^{(s)}|y)}$$

- Monte Carlo methods which can sample from $p(\theta^{(s)}|y)$ using only $q(\theta^{(s)}|y)$

$$E\left(f(\theta) \mid y\right) \approx \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)})$$

Approximating expectations

A few different approaches of computation in the Bayesian context:

- Conjugate priors and analytic solutions (Ch 1-5)
- Grid integration and other quadrature rules (Ch 3, 10)
- Independent Monte Carlo, rejection and importance sampling (Ch 10)
- Markov Chain Monte Carlo (Ch 11-12)
- Other distributional approximations (Laplace, variational Bayes, expectation propagation) (Ch 4, 13)

*Numerical accuracy

- Floating point presentation of numbers. e.g. with 64bits, the closest value to zero is $\approx 2.2 \cdot 10^{-308}$
- Joint densities with lots of product components result in underflow, e.g.
  `prod(dnorm(rnorm(600)))` $\rightarrow$ `0`.
- Use log densities to avoid over- and underflows in floating point presentation, e.g.
  `sum(dnorm(rnorm(600),log=TRUE))` $\rightarrow$ `-847.3`,
  so that you can handle a lot more observations.
- Compute `exp` as late as possible.

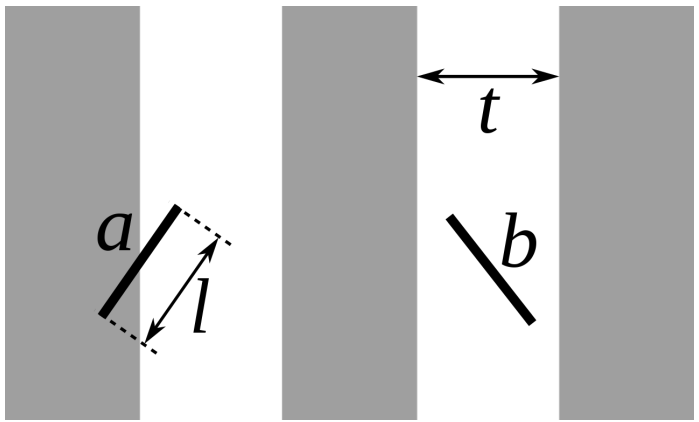
Monte Carlo methods

Monte Carlo samples can be used, e.g. to compute means, deviations, quantiles; to draw histograms; to marginalize posterior distributions, etc.

- The idea existed and was practiced way before computers (Buffon's Needle, 1777)
- The term "Monte Carlo method" was proposed by Metropolis, von Neumann and Ulam towards the end of 1940s
- 1990s: the Monte Carlo revolution in Bayesian computation
- Idea: simulate draws from the target distribution, and treat these draws as observations. A collection of draws is thus a sample.

Buffon's Needle

[Wikipedia] Suppose we have a floor made of parallel strips of wood, each the same width t. We drop a needle of length l onto the floor. What is the probability that the needle will lie across a line between two strips? – *Georges-Louis Leclerc, Comte de Buffon, 1777*



See https://mste.illinois.edu/activity/buffon/ for an illustration of the analysis.

How many simulation draws are needed?

If draws are **independent**, use usual methods to estimate the uncertainty due to a finite number of observations.

Posterior expectation

$$\mathbb{E}(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \theta^{(s)}$$

if S is big and $\theta^{(s)}$ are independent, way may assume that the distribution of the expectation approaches normal distribution with variance σ_θ^2 / S (asymptotic normality).

How many simulation draws are needed?

Posterior expectation

$$\mathbb{E}(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \theta^{(s)}$$

- Total variance of the estimator is the sum of the epistemic uncertainty in the posterior, and the uncertainty due to using finite number of Monte Carlo draws

$$\sigma_\theta^2 + \sigma_\theta^2 / S = \sigma_\theta^2 (1 + 1/S)$$

- For example, if $S = 100$, deviation increases by $\sqrt{1 + 1/S} = 1.005$, i.e. Monte Carlo error is very small.
- *Counter examples for asymptotic normality (Chapter 4)

How many simulation draws are needed?

Posterior probability

$$p(\theta \in A) \approx \frac{1}{S} \sum_{I} I(\theta^{(s)} \in A),$$

where $I(\theta^{(s)} \in A) = 1$ if $\theta^{(s)} \in A$.

- $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
  - $var(I(\cdot)) = p(1-p)$
  - standard deviation of the estimate is $\sqrt{p(1-p)/S}$
- For example, if $S = 100$ and $p \approx 0.5$, $\sqrt{p(1-p)/S} = 0.05$, i.e. accuracy is about 5% units. Thus, $S = 2500$ draws are needed for 1% unit accuracy.

To estimate small probabilities, a large number of draws is needed

- To be able to estimate p, need to get draws with $\theta^{(I)} \in A$, which in expectation requires $S \gg 1/p$.

How many simulation draws are needed?

Note. The number of independent draws needed does **not** depend on the number of dimensions. However, it may be difficult to obtain independent draws in a high dimensional case.

Markov chain Monte Carlo produces **dependent** draws, and requires additional work to estimate the **effective sample size**.

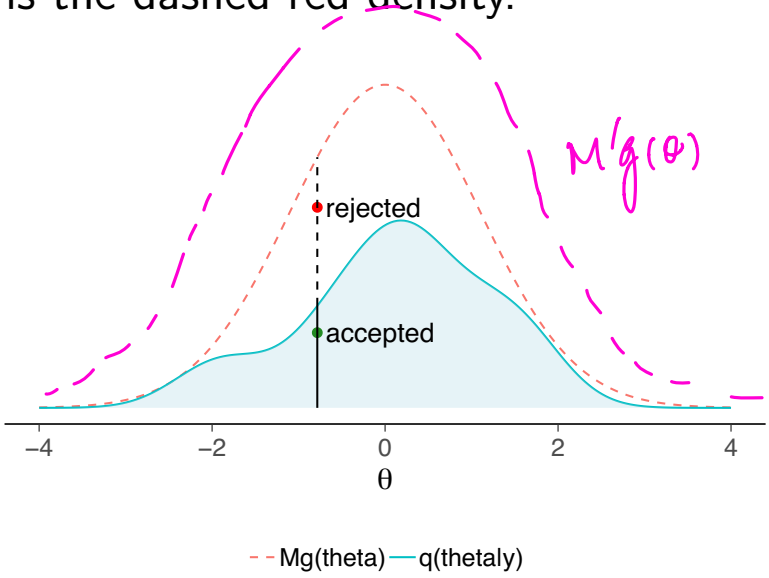
- Variance reduction methods

Rejection sampling

Choose a proposal distribution $g(\cdot)$ such that it can form an envelope over the (unormalized) target distribution $q(\cdot) = cp(\cdot)$, i.e. there exists a *covering constant* $M < \infty$ s.t. for all θ,

$$Mg(\theta) \geq q(\theta).$$

Below, the unormalized target $q(\cdot)$ is the blue density. Scaled proposal $Mg(\cdot)$ is the dashed red density.

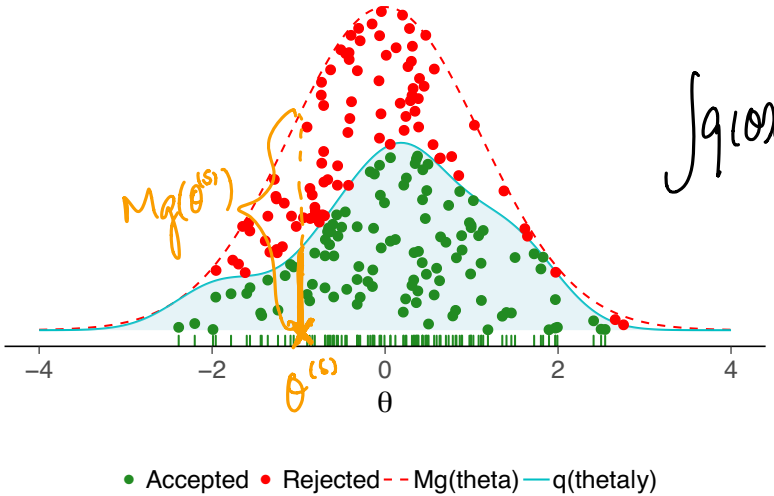


Rejection sampling

(von Neumann, 1951). At step $s = 1, \ldots, S$:

- Draw a sample $\theta^{(s)} \sim g(\cdot)$;
- Accept $\theta^{(s)}$ with probability $r = \frac{q(\theta^{(s)})}{Mg(\theta^{(s)})}$, otherwise go back to the previous step.

The accepted samples $\theta^{(s)}$ follow the target distribution $p(\cdot)$.



Rejection sampling: proof

Let $I = \begin{cases} 1 & \text{if } \theta^{(s)} \sim g \text{ is accepted} \\ 0 & \text{o/w} \end{cases}$

$$\Pr(I=1) = \int \Pr(I=1|\theta) g(\theta) d\theta = \int \frac{cp(\theta)}{Mg(\theta)} g(\theta) d\theta$$

$$= \frac{c}{M}$$

So $$\Pr(\theta | I=1) = \frac{\frac{cp(\theta)}{Mg(\theta)} g(\theta)}{\Pr(I=1)} = p(\theta).$$

Rejection sampling: Efficiency

- The number of accepted draws (out of S) is the effective sample size.
- The covering constant M reflects the *expected* number of operations needed to obtain one draw, so the key is to find a good proposal g with a small M.
- Selection of good proposal gets very difficult when the number of dimensions increase.

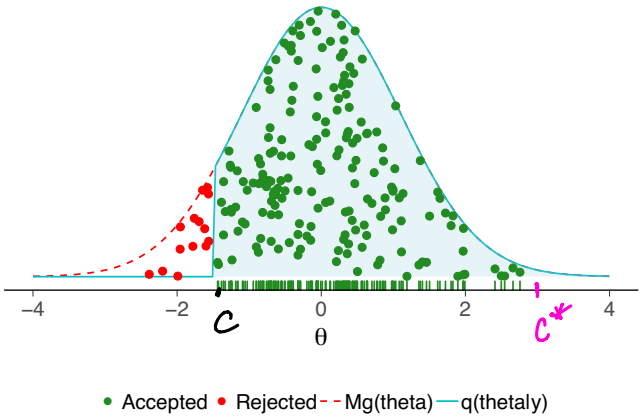
Example: Truncated Gaussian

The (unormalized) target is

$$q(\theta) = \phi(\theta)\,\mathbf{1}\{\theta > c\}$$

where ϕ is the standard Normal density.

- $c \leq 0$: choose $g(\theta) = \phi(\theta)$. The covering constant $M = 1$. Efficiency is at least 50% (depending on c).



- $c > 0$, especially $c \gg 0$: this choice can be inefficient.

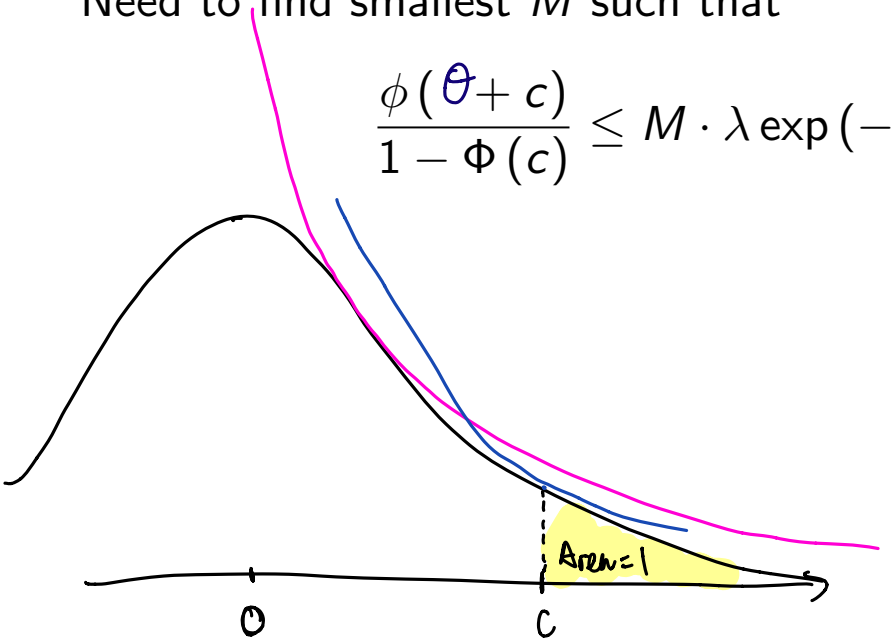
Example: Truncated Gaussian

Consider proposal distribution in the form of the exponential density (with parameter λ as something we can choose):

$$g(\theta) = \lambda \exp(-\lambda\theta).$$

Need to find smallest M such that

$$\frac{\phi(\theta + c)}{1 - \Phi(c)} \leq M \cdot \lambda \exp(-\lambda\theta), \ \forall\theta > 0.$$



Example: Truncated Gaussian

$$\frac{\phi(\theta + c)}{1 - \Phi(c)} \leq M \cdot \lambda \exp(-\lambda\theta), \ \forall \theta > 0$$

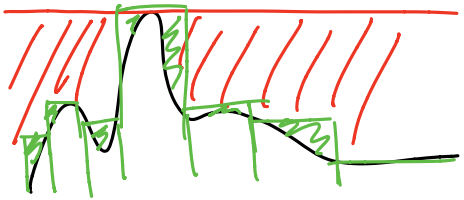
gives the minimal M^*:

$$M^* = \frac{\exp\left((\lambda^2 - 2\lambda c)/2\right)}{\sqrt{2\pi}\,\lambda\,(1 - \Phi(c))}$$

The best choice λ^* that meets this minimum rejection rate:

$$\lambda^* = \left(c + \sqrt{c^2 + 4}\right)/2$$

Variance reduction methods



- Stratified sampling: break Θ into k disjoint regions such that the target integral within the subregion is relatively homogeneous;
- Control variates: use a control variate that is correlated with the sample θ, to produce a better estimate (similar to regression adjustment);
- **Antithetic variates**
- **Rao-Blackwellization**

Antithetic variates

(Hammersley and Morton, 1956) This is a method to produce negatively correlated samples. Suppose $U \sim Unif\,(0,1)$, and we draw $X = F^{-1}(U)$, i.e. via PIT, to obtain $X \sim F$.

Then, using the same draw of U,

$$X' = F^{-1}(1-U)$$

also follows the same distribution $X' \sim F$.

Antithetic variates

More generally, if h is a monotone function, then

$$(h(u) - h(v))(h(1-u) - h(1-v)) \leq 0$$

for every $u, v \in [0,1]$. Thus for two i.i.d. Uniform r.v.s U and V, we have that

$$\mathbb{E}(h(U) - h(V))(h(1-U) - h(1-V)) = Cov(X, X') \leq 0$$

where $X = h(U)$ and $X' = h(1-U)$. Therefore

$$var\left(\frac{X + X'}{2}\right) \leq \frac{var(X)}{2}$$

* coupling

implying that using the pair (X, X') is better than using two independent Monte Carlo draws for estimating $\mathbb{E}(X)$.

Rao-Blackwellization

In MC computation, **always carry out analytical computation as much as possible**.

Suppose you have S independent draws $x^{(s)} \sim p(\cdot)$, and want to estimate $E(f(x))$. The MC estimate is

$$\hat{f} = \frac{1}{S} \sum_s f\left(x^{(s)}\right).$$

Now suppose $x = (x_1, x_2)$ has two parts, and you know the conditional expectation

$$E\left(f\left(x\right) \mid x_2\right).$$

Consider the alternative MC estimator

$$\tilde{f} = \frac{1}{S} \sum_s E\left(f\left(x\right) \mid x_2^{(s)}\right).$$

Rao-Blackwellization

$$\hat{f} = \frac{1}{S}\sum_s f\left(x^{(s)}\right), \qquad \tilde{f} = \frac{1}{S}\sum_s E\left(f(x) \mid x_2^{(s)}\right).$$

Both $\hat{f}$ and $\tilde{f}$ are unbiased:

$$E_p(f(x)) = E\left(E(f(x) \mid x_2)\right)$$

Thus, estimator $\tilde{f}$ should be preferred since $var\left(\hat{f}\right) \geq var\left(\tilde{f}\right)$:

$$Var_p(f(x)) = E\left(Var(f(x) \mid x_2)\right) + Var\left(E(f(x) \mid x_2)\right)$$

$$\geq 0.$$

$$Var(\hat{f}) = \frac{Var(f(x))}{S} \geq \frac{Var\left(E(f(x) \mid x_2)\right)}{S} = Var(\tilde{f})$$

Importance Sampling

(Marshall 1956) Importance Sampling focus on regions "of importance" to save computational resources. It is related to rejection sampling and a precursor to the Metropolis algorithm.

Suppose we want to estimate the posterior expectation of $f(\theta)$. The unormalized target is $q(\theta) = cp(\theta)$ for some $c < \infty$, possibly unknown.

Write

$$E\left(f(\theta) \mid y\right) = \int f(\theta)p(\theta)d\theta = \frac{\int f(\theta)q(\theta)d\theta}{\int q(\theta)d\theta} = \frac{\int f(\theta)\frac{q(\theta)}{g(\theta)}g\left(\theta\right)d\theta}{\int \frac{q(\theta)}{g(\theta)}g\left(\theta\right)d\theta}.$$

Importance Sampling

$$E\left(f(\theta) \mid y\right) = \int f(\theta)p(\theta)d\theta = \frac{\int f(\theta)q(\theta)d\theta}{\int q(\theta)d\theta} = \frac{\int f(\theta)\frac{q(\theta)}{g(\theta)}g\left(\theta\right)d\theta}{\int \frac{q(\theta)}{g(\theta)}g\left(\theta\right)d\theta}$$

can be estimated using S draws $\theta^{(1)}, \ldots, \theta^{(S)}$ from $g\left(\cdot\right)$ by the expression

$$\hat{f} = \frac{\sum_s w^{(s)} f(\theta^{(s)})}{\sum_s w^{(s)}},$$

where

$$w^{(s)} = \frac{q(\theta^{(s)})}{g(\theta^{(s)})} \propto \frac{p(\theta^{(s)})}{g(\theta^{(s)})}$$

are the **importance weights**. See `R` for an illustration.

Problem!
$w^{(s)} \gg 1$

Importance Sampling

- In general,

$$\hat{f} = \frac{\sum_s w^{(s)} f(\theta^{(s)})}{\sum_s w^{(s)}}$$

  is **biased** for $E\left(f(\theta) \mid y\right)$, due to the estimated weights in the denominator.
- If the ratio $p(\cdot)/g(\cdot)$ is exactly known, then we have an alternative **unbiased** estimate

$$\tilde{f} = \frac{1}{S} \sum_s w^{(s)} f(\theta^{(s)}).$$

  However, $\hat{f}$ often has a smaller MSE than $\tilde{f}$.

Choice of IS proposal

The importance weights

$$w^{(s)} = \frac{q(\theta^{(s)})}{g(\theta^{(s)})} \propto \frac{p(\theta^{(s)})}{g(\theta^{(s)})}$$

are used to correct for the "bias" in drawing the θ's from the proposal $g(\cdot)$ as opposed to the target $p(\cdot)$.

- A good choice for the proposal $g(\cdot)$ is one that's close in shape to $f(\cdot)p(\cdot)$.
- Unlike rejection sampling, the proposal $g(\cdot)$ need not cover the target $p(\cdot)$. However, it is desirable that it has a longer tail than $p(\cdot)$, to avoid importance weights with large variance.

Effective sample size

Suppose we can compute the normalized target, and the naturally normalized importance weights can be obtained as

$$w_0^{(s)} = \frac{p(\theta^{(s)})}{g(\theta^{(s)})}.$$

Note that here, $E_g\left(w_0\left(\theta\right)\right) = 1$.

The **effective sample size** (ess) of S independent random θ samples generated from target g is defined as

$$ess\left(S\right) = \frac{S}{1 + var_g\left(w_0\left(\theta\right)\right)}.$$

Effective sample size

If only the unormalized q is known, we cannot compute w_0, only

$$w^{(s)} = \frac{q(\theta^{(s)})}{g(\theta^{(s)})}.$$

The variance of the normalized weights $var_g\left(w_0\left(\theta\right)\right)$, if finite, needs to be estimated with the **coefficient of variation** of the unormalized weights. That is,

$$\hat{ess}\left(S\right) = \frac{S}{1 + cv^2\left(w\right)},$$

where $\bar{w} = \frac{1}{S}\sum_s w^{(s)}$, and

$$cv^2\left(w\right) = \frac{\sum_s \left(w^{(s)} - \bar{w}\right)^2}{\left(S-1\right)\bar{w}^2}.$$

Effective sample size

Exercise. Compare with the *ess* approximation in the textbook, (10.4) on page 266:

$$\tilde{ess}(S) = \frac{1}{\sum_s \left(\tilde{w}^{(s)}\right)^2},$$

where

$$\tilde{w}^{(s)} = \frac{w^{(s)}}{\sum_s w^{(s)}}$$

Note 1: The textbook has a typo here. There is no S multiplier in the expression of $\tilde{w}$.
Note 2: The approximated effective sample size is small if there are a few extremely large weights. These few but large weights make the estimate itself very noisy, so take it with a grain of salt!