# MSA101/MVE187 2021 Lecture 4
# Inference by simulation: Monte Carlo Integration
# Basic simulation methods
# Rejection sampling

Petter Mostad

Chalmers University

September 9, 2021

# Review and overview

- We have looked at the Bayesian paradigm, conjugacy, some fundamental properties.
- Our examples have been super-simple applications.
- In many realistic cases the relationship between $y_{pred}$ and $y_{data}$ needs a complicated model with many parameters to describe it: In other words, a high-dimensional $\theta$.
- Then, how to compute? A possibility is
  - to generate an (approximate) random sample from $\pi(\theta \mid y_{data})$.
  - Then use that sample to approximate $\pi(y_{pred} \mid y_{data}) = \int \pi(y_{pred} \mid \theta)\pi(\theta \mid y_{data})\, d\theta$.
- Today, we look at how to do the second step above.
- We also start looking at how to generate random samples.

# Monte Carlo Integration

Assume $\theta_1, \theta_2, \ldots, \theta_N$ is a random sample from $\pi(\theta \mid y)$.

- $\Pr(\theta > z) \approx \frac{\#\ \theta_i\text{'s above } z}{N}$.
- We can rewrite this in a fancy way as

$$E_{\theta|y}(I(\theta > z)) = \int I(\theta > z)\pi(\theta \mid y)\, d\theta \approx \frac{1}{N}\sum_{i=1}^{N} I(\theta_i > z).$$

- More generally (assuming the expectation exists)

$$E_{\theta|y}(f(\theta)) = \int f(\theta)\pi(\theta \mid y)\, d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i).$$

- Formally, according to the Strong Law of large numbers,

$$\Pr\left(\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(\theta_i) = E(f(\theta))\right) = 1$$

  where the expectation is taken over a distribution from which $\theta_1, \ldots, \theta_N$ is a random sample.

# Using Monte Carlo integration for predictions

- Example: To approximate a probability $\Pr(y_{pred} > z \mid y_{data}) = \int \Pr(y_{pred} > z \mid \theta)\, \pi(\theta \mid y_{data})\, d\theta$
  - Generate $\theta_1, \ldots, \theta_N$ from the posterior for $\theta$ given $y_{data}$.
  - Use as approximation $\frac{1}{N} \sum_{i=1}^{N} \Pr(y_{pred} > z \mid \theta_i)$ .
- Example: If $\theta = (\alpha, \beta, \gamma)$ is the parameter vector, what is the posterior probability that $\alpha > \beta^2$?
- Solution: We generate a set of vectors $\theta_1, \ldots, \theta_N$ from the posterior for $\theta$ given $y_{data}$. Then:
- Approximate $\Pr\left(\alpha > \beta^2 \mid y_{data}\right)$ with

$$\frac{1}{N} \sum_{i=1}^{N} I(\alpha_i > \beta_i^2)$$

  where $\theta_i = (\alpha_i, \beta_i, \gamma_i)$ .
- **Presentation break for computations by hand**

# Simulation of predicted values

- Approximating the value of $\Pr(y > z \mid y_{data})$ in two ways:
- Alternative 1 (as above):
  - Simulate $\theta_1, \ldots, \theta_N$ from the posterior of $\theta$ given $y_{data}$.
  - Compute

$$\frac{1}{N} \sum_{i=1}^{N} \Pr(y > z \mid \theta_i)$$

- Alternative 2:
  - Use $\pi(y \mid \theta)$ to simulate posterior values for $y$ together with posterior values for $\theta$: We get $(\theta_1, y_1), (\theta_2, y_2), \ldots, (\theta_N, y_N)$.
  - Compute

$$\frac{1}{N} \sum_{i=1}^{N} I(y_i > z)$$

- **Presentation break for computations by hand**

# Example: Approximating quantiles by simulation

- A 95% *credibility interval* for a random variable $X$ is an interval so that the probabiliy that $X$ is in the interval is 95%.
- In Bayesian statistics, a posterior credibility interval for a variable $y$ may be used to describe the posterior uncertainty in $y$.
- A way to approximate a 90% posterior credibility interval for $y$:
  - Simulate a posterior sample $y_1, y_2, \dots, y_N$ as above.
  - Order by size to find the 5th and 95th empirical quantiles of $y_1, \dots, y_N$. (In R, use `quantile(y, c(0.05, 0.95))`.)
- **Presentation break for computations by hand**

# Accuracy of Monte Carlo integration

- Assume $\theta_1, \theta_2, \ldots, \theta_N$ is a random sample from $\pi(\theta \mid y)$. The Central Limit Theorem (CLT) states that, approximately for large $N$,

$$\frac{1}{N}\sum_{i=1}^{N} f(\theta_i) \sim \text{Normal}\left(\mathrm{E}_{\theta|y}(f(\theta)), \frac{\mathrm{Var}_{\theta_|y}(f(\theta))}{N}\right)$$

  as long as the first two moments of $f(\theta)$ exist.
- Transferring to a Bayesian setting (and using a flat prior) we get that, after sampling $\theta_1, \ldots, \theta_N$, an approximate 95% credibility interval for $\mathrm{E}_{\theta|y}(f(\theta))$ is

$$\frac{1}{N}\sum_{i=1}^{N} f(\theta_i) \pm 1.96\frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}_{\theta|y}(f(\theta))}.$$

- If we write $\overline{f(\theta)} = \sum_{i=1}^{N} f(\theta_i)/N$ we may approximate

$$\mathrm{Var}_{\theta|y}(f(\theta)) \approx s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(f(\theta_i) - \overline{f(\theta)}\right)^2.$$

# Example: Estimating a proportion

- Let's say we want to approximate the proportion of an (posterior) random variable that is below $z$. For a sample of size $N$, we find that $r$ are below $z$.
- Plugging into the formula above gives the estimate

  $$\frac{r}{N}$$

  together with the 95% credibility interval

  $$\left[\frac{r}{N} - 1.96\frac{s}{\sqrt{N}}, \frac{r}{N} + 1.96\frac{s}{\sqrt{N}}\right]$$

  where

  $$s^2 = \frac{r(N-r)}{N(N-1)}$$

- **Presentation break for computations by hand**

# Bayesian inference using simulation

- We want to do Bayesian inference by
  - simulating a sample $\theta_1, \ldots, \theta_N$ from the posterior of $\theta$ given $y_{data}$.
  - making predictions based on this posterior sample.
- The second part has basically been covered above. The first part will take up half of the rest of the course.
- We use Bayes formula to find the posterior density:

$$\pi(\theta \mid y_{data}) = \frac{\pi(y_{data} \mid \theta)\pi(\theta)}{\pi(y_{data})} \propto_\theta \pi(y_{data} \mid \theta)\pi(\theta)$$

- In many cases we have formulas for the likelihood $\pi(y_{data} \mid \theta)$ and the prior $\pi(\theta)$ but *not* for $\pi(y_{data})$.
- Solution: We develop methods that produce an (approximate) sample based only on a formula for the density multiplied by an unknown constant.
- First, we start with the basics of computer simulation of random variables.

# Simulation from a uniform distribution

- Simulation from Uniform$[0,1]$ is the basis of all computer based simulation.
- What does it mean that $x_1, \ldots, x_n \sim$ Uniform$[0,1]$ is "random"? A possible interpretation: We have no way to predict the coming numbers; the best guess for their distribution is Uniform$[0,1]$.
- The computer uses a deterministic function applied to a seed ("pseudo-random"). The seed can be set (in R with `set.seed(...)`) or is taken from the computer clock.
- It should be in practice impossible to apply any kind of visualiation or compute any kind of statistic which has properties other than those predicted when the sequence $x_1, \ldots, x_n$ is *iid* Uniform$[0,1]$.

# Simulating from discrete distributions

- If $X$ is a random variable on a finite set of real numbers, the cumulative distribution can be computed in a vector. $X$ can be simulated by comparing a uniform random variable $U$ to the numbers in this vector. Example: Binomial distribution.
- **Presentation break for computations by hand**
- If $X$ is a random variable on a countable set of real numbers, one can use a list of the probabilities of the most probable outcomes, and expand this list as needed, if extreme values are simulated in a uniform distribution. Example: The Poisson distribution.

# The inverse transform

- Let $X$ be a random variable with invertible cumulative distribution function $F(x)$. If $U \sim \text{Uniform}[0,1]$, then $F^{-1}(U)$ is a random sample from X.
- Proof:

  $$\Pr(F^{-1}(U) \leq \alpha) = \Pr(F(F^{-1}(U)) \leq F(\alpha)) = \Pr(U \leq F(\alpha)) = F(\alpha)$$

- Example: The exponential distribution $\text{Exp}(\lambda)$ has density $\pi(X) = \lambda \exp(-x\lambda)$ and cumulative distribution

  $$F(x) = 1 - \exp(-\lambda x)$$

  $F(x) = u$ gives $F^{-1}(u) = -\log(1-u)/\lambda$. As $1-u$ is uniform, we can simulate with

  $$-log(u)/\lambda$$

- **Presentation break for computations by hand**

# The inverse transform, cont.

- Example: Logistic distribution. Best defined by defining its cumulative distribution (for standard logistic distribution):

  $$F(x) = 1/(1 + \exp(-x))$$

  Easy to invert. The distribution can be adjusted with changing the mean and the scale.
- Example: Cauchy distribution. Density:

  $$\pi(x) = 1/(\pi(1 + x^2)).$$

  The cumulative distribution is

  $$F(x) = 1/2 + 1/\pi \arctan(x)$$

  Easy to invert.

# Transforming samples

- Example: One can prove that, if $x_1, \ldots, x_n$ is a random sample from Exp(1) then
$$\frac{1}{\beta}\sum_{i=1}^{n} x_i \sim \text{Gamma}(n, \beta)$$
- Example: One can prove that, if $x_1, \ldots, x_{a+b}$ is a random sample from Exp(1) then
$$\frac{\sum_{i=1}^{a} x_i}{\sum_{i=1}^{a+b} x_i} \sim \text{Beta}(a, b).$$
- Example: One can prove that, if $u_1, u_2$ is a random sample from Uniform$[0, 1]$, then
$$\left(\sqrt{-2\log(u_1)}\cos(2\pi u_2), \sqrt{-2\log(u_1)}\sin(2\pi u_2)\right)$$
is a random sample from the bivariate distribution Normal $\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$.

# Transformation of random variables

- Recall from basic probability theory: If $f(x)$ is a density function, and $x = h(y)$ is a monotone transformation, then the density function for $y$ is
$$f(h(y))|h'(y)|$$
- If we apply the INVERSE of $h$ on a variable with known density, we get the density of the resulting variable using the formula above.
- Example application: The non-informative prior for the precision $\tau$ of a Normal distribution is the improper distribution with "density" $\pi(\tau) \propto 1/\tau$. We have that $\tau = h(\sigma^2) = 1/\sigma^2$. With $h(x) = 1/x$ we get that $h'(x) = -1/x^2$. Thus the corresponding non-informative prior for the variance $\sigma^2$ of a normal distribution is given as
$$\pi(\sigma^2) \propto \frac{1}{1/\sigma^2}\left|-\frac{1}{(\sigma^2)^2}\right| = \frac{1}{\sigma^2}.$$

# Transformation of multivariate random variables

- If $x$ is a vector, if $f(x)$ is a multivariate density function, and if $x = h(y)$ is a bijective differentiable transformation, then the multivariate density function for $y$ is

$$f(h(y))|J(y)|$$

  where $|J(y)|$ is the determinant of the Jacobian matrix for the vector function h(y).
- One application of this is in the proof of the formula used above to sample from the bivariate normal distribution.

# Rejection sampling

- Sometimes we cannot easily simulate from a density $f(x)$, (the "target density") but we *can* simulate from an "instrumental" density $g(x)$ that approximates $f(x)$.
- If we can find a constant $M$ such that $f(x)/g(x) \leq M$ for all $x$ in the support of $g$ and $f(x) = 0$ outside this support, we can use *rejection sampling* to sample from $f$:
  - Sample $x$ from the distribution with density $g(x)$.
  - Draw $u$ uniformly on $[0, 1]$.
  - If $u \cdot M \cdot g(x) \leq f(x)$ accept $x$ as a sample, otherwise reject $x$ and start again.
- **Presentation break for computations by hand**

# Rejection sampling, cont.

- We may in fact do this with $f(x) = C\pi(x)$ where $\pi(x)$ is the actual density and $C$ is unknown: It is still a valid method!
- When $f(x)$ integrates to 1, the acceptance rate is $1/M$, so we want to use a small $M$.
- When $f(x)$ does not integrate to 1, the integral can be approximated as the acceptance rate multiplied by $M$.
- NOTE: Applicable for $x$ of any dimension!
- Example: Random variables with picewise log-concave densities can be simulated with this method.
- **Presentation break for computations by hand**

# Simulating from the multivariate normal

- Recall that $x \sim \text{Normal}_k(\mu, \Sigma)$ if

$$\pi(x) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right)$$

- NOTE: If $x_1, \ldots, x_k$ are i.i.d Normal$(0, 1)$ then $x = (x_1, \ldots, x_n)^t \sim \text{Normal}_k(0, I)$.
- If $x \sim \text{Normal}_k(0, I)$ then $Ax \sim \text{Normal}(0, AA^t)$.
- THUS: To simulate from Normal$(\mu, \Sigma)$:
  - Simulate $k$ independent standard normal random variables into a vector $x$.
  - Compute the (lower triangular) Choleski decomposition $S$ of $\Sigma$: We then have that $\Sigma = SS^t$.
  - Compute $Sx + \mu$: It is multivariate normal, and has the right expectation and covariance matrix.

# Simulating from a marginal distribution

- Generally: If you have a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ from a joint distribution of $x$ and $y$, then $x_1, x_2, \ldots, x_n$ is a sample from the marginal distribution of $x$.
- Simple application: If $\tau \sim \text{Gamma}(k/2, 1/2)$ and $x \mid \tau \sim \text{Normal}(0, 1/\tau)$, then the marginal distribution of $x$ is a Student t-distribution with $k$ degrees of freedom. To simulate:
  - Draw $\tau$ from $\text{Gamma}(k/2, 1/2)$.
  - Then draw $x$ from $\text{Normal}(0, 1/\tau)$.
- Much more generally: To simulate for example from the predictive distribution in a Bayesian model, simulate from the joint distribution with density $\pi(y, \theta)$. Then take the coordinates of the sample pertaining to $y$.