MSA101/MVE187 2020 Lecture 4.1
Inference by simulation
Monte Carlo Integration

Petter Mostad

Chalmers University

September 4, 2020

Overview of Bayesian inference so far

- A stochastic *model* (joint probability density) for all variables is constructed.
  - Variables represent **data** (x), **unknown parmeters** (θ), and **values you want to predict** (x_{NEW}).
  - Generally, the joint model is presented as $\pi(x, x_{NEW}, \theta) = \pi(x, x_{NEW} \mid \theta)\pi(\theta)$, where $\pi(\theta)$ the *prior*. Often we have independence given the parameter θ, so that $\pi(x, x_{NEW} \mid \theta) = \pi(x \mid \theta)\pi(x_{NEW} \mid \theta)$ where $\pi(x \mid \theta)$ is the *likelihood*.
- Generally, the goal is to make posterior predictions of x_{NEW}, so to describe the posterior predictive $\pi(x_{NEW} \mid x)$. Either one may do this directly, or first find the posterior $\pi(\theta \mid x)$ and then the posterior predictive $\pi(x_{NEW} \mid x) = \int \pi(x_{NEW} \mid \theta)\pi(\theta \mid x)\, d\theta$.
- In the simplest models, the posterior and posterior predictive can be computed analytically using conjugacy.
- When the total number of unknown variables in the model is small (1-3?) you may use numerical discretization to find the posterior.
- In all other cases, we need other methods.

The next idea: Bayesian inference using simulation

Assume we have a model $\pi(x, x_{NEW}, \theta)$. Do inference in two steps:

- Generate $\theta_1, \theta_2, \ldots, \theta_N$ as an (approximate) sample from the posterior $\pi(\theta \mid x)$.
- Use these θ values to obtain information about the posterior of θ or the posterior predictive for x_{NEW}:
  - The posterior distribution of θ can be investigated by approximating it with $\theta_1, \theta_2, \ldots, \theta_N$.
  - The predictions for x_{NEW} can be obtain by averaging the predictions $\pi(x_{NEW} \mid \theta)$ over $\theta_1, \ldots, \theta_N$.

Monte Carlo Integration

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

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

$$\mathrm{E}_{\theta|x}(I(\theta > z)) = \int I(\theta > z)\pi(\theta \mid x)\, dx \approx \frac{1}{N}\sum_{i=1}^{N} I(\theta_i > z).$$

- More generally, $\mathrm{E}_{\theta|x}(f(\theta)) \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)$.
- Formally, according to the Strong Law of large numbers,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(\theta_i) = \mathrm{E}_{\theta|x}(f(\theta)).$$

- Making predictions:

$$\begin{aligned} \pi(x_{NEW} \mid x) &= \int \pi(x_{NEW} \mid \theta)\pi(\theta \mid x)\, d\theta = \mathrm{E}_{\theta|x}(\pi(x_{new} \mid \theta)) \\ &\approx \frac{1}{N}\sum_{i=1}^{N} \pi(x_{NEW} \mid \theta_i) \end{aligned}$$

Accuracy of Monte Carlo integration

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

- The Central Limit Theorem (CLT) states that, approximately,

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

  as long as the first two moments of $f(\theta)$ exist.
- Thus a traditional 95% approximate confidence interval for $\text{E}_{\theta|x}(f(\theta))$ is

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

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

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

  we get the approximate confidence interval $\frac{1}{N}\sum_{i=1}^{N} f(\theta_i) \pm 1.96\frac{s}{\sqrt{N}}$.

MSA101/MVE187 2020 Lecture 4.2 Examples

Petter Mostad

Chalmers University

September 4, 2020

Example: Estimating a proportion

- Assume $f(\theta) = I(g(\theta) \leq \alpha)$ so that we want to estimate the posterior probability $p = \Pr(g(\theta) \leq \alpha)$ that $g(\theta)$ is below α.
- Following the above, the estimate $\hat{p} = \sum_{i=1}^{N} f(\theta_i)/N$ consists of counting the proportion of the $g(\theta_i)$ that is below α.
- We get

$$\mathrm{Var}(I(g(\theta \leq \alpha))) = \mathrm{E}(I(g(\theta \leq \alpha))) - \mathrm{E}(I(g(\theta \leq \alpha)))^2 = p - p^2.$$

Thus the *relative variability* of an estimate can be estimated as

$$\frac{\sqrt{\mathrm{Var}(f(\theta))}}{\mathrm{E}(f(\theta))} \approx \frac{\sqrt{\hat{p} - \hat{p}^2}}{\hat{p}} = \sqrt{1/\hat{p} - 1}.$$

which is large when $\hat{p}$ is small.

- In other words: Estimating a tail quantile for a probability distribution by counting the number of times sampled values are in the tail may not give the desired accuracy.

Example: Approximating quantiles by simulation

To compute an approximate interval containing, e.g., 90% of the probability for a random variable X:

- Simlulate $x_1, \ldots, x_n$ from X.
- Order them by size and fiind the 5'th and 95'th empirical percentile.
- In R, use, e.g., `quantile(..)`.

Example: Our old friend the Beta-Binomial conjugacy

- 6 successes in 19 trials observed. Probability of success p has a flat prior on $[0, 1]$. What is the probability of 4 or more successes in 7 new trials?
- y: number of successes in first trials. y_n: number of successes in new trials. Stochastic model:

$$\pi(y, y_n, p) = \pi(y \mid p)\pi(y_n \mid p)\pi(p) = \text{Binomial}(y; 19, p) \cdot \text{Binomial}(y_n; 7, p)$$

- The prior for p can be expressed as $\text{Beta}(1, 1)$ and because of conjugacy the posterior for p then becomes $\text{Beta}(7, 14)$.
- Three alternative ways to compute with R the answer to the question above:
  - Use simulation from the posterior for p.
  - Use conjugacy to to compute the posterior predictive for y_n given y.
  - Use discretization.

MSA101/MVE187 2020 Lecture 4.3
Simulation of random variables

Petter Mostad

Chalmers University

September 4, 2020

Simulation of random variables

- As discussed above, we would like to obtain an (approximate) sample $\theta_1, \ldots, \theta_N$ from the posterior $\pi(\theta \mid x)$. How do we do it?
- θ may be high dimensional, and the posteior $\pi(\theta \mid x)$ may be a complicated function, maybe known only up to a constant if it is computed using $\pi(\theta \mid x) \propto_\theta \pi(x \mid \theta)\pi(\theta)$.
- In this course we will mainly simulate with Markov chain Monte Carlo (MCMC) in such cases.
- But first, we go through some simpler simulation methods:
  - Some basic simulation methods
  - Rejection sampling
  - Slice sampling (in later lecture)

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.
- 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.
- Note:

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

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

$$-1/\lambda log(u)$$

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

$$\beta \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \beta)$$

- Example: One can prove that, if $X_1, \ldots, X_n$ 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 τ 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 σ^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.

MSA101/MVE187 2020 Lecture 4.4 Rejection sampling. Sampling from the multivariate normal.

Petter Mostad

Chalmers University

September 4, 2020

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 (and if f and g have the same support), we can use *rejection sampling* to sample from f:
  - Sample X using $g(x)$.
  - 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.

Rejection sampling, cont.

- NOTE: Applicable in any dimension.
- The acceptance rate is $1/M$ (when both f and g are actual densities), so we want to use a small M.
- NOTE: We may in fact do this with $f(x)$ and $g(x)$ equal to the densities up to a constant, still a valid method!
- NOTE: When $g(x)$ integrates to 1, the integral of $f(x)$ can be approximated as the acceptance rate multiplied by M.
- Example: Random variables with log-concave densities can be simulated with this method.

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(μ, Σ):
  - Simulate k independent standard normal random variables into a vector x.
  - Compute the (lower triangular) Choleski decomposition S of Σ: 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 τ 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 for x_{NEW} in a Bayesian model, simulate from the joint distribution with density $\pi(x_{NEW}, \theta \mid x)$, where x is the data and θ is the parameters. Then take the coordinates of the sample pertaining to x_{NEW}.