MSA101/MVE187 2022 Lecture 4
Inference by simulation. Monte Carlo Integration
Basic simulation methods
Rejection sampling. Priors

Petter Mostad

Chalmers University

September 7, 2022

Predictions using simulation

- We may want to make predictions by simulating from marginal distributions, e.g.,

$$\begin{aligned} \pi(y) &= \int \pi(y \mid \theta)\pi(\theta)\, d\theta \\ \pi(y \mid y_{\text{data}}) &= \int \pi(y \mid \theta)\pi(\theta \mid y_{\text{data}})\, d\theta \end{aligned}$$

- Generate a sample $(\theta_1, y_1), \ldots, (\theta_N, y_n)$ from the joint density!
- Generate the sample by first simulating $\theta_1, \ldots, \theta_N$ from $\pi(\theta)$ (or $\pi(\theta \mid y_{\text{data}})$) and then simulate y_i from $\pi(y \mid \theta_i)$ for $i = 1, \ldots, N$.
- Then $y_1, \ldots, y_N$ is a sample from the marginal.

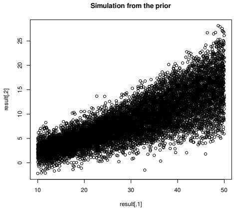
Example: Simulating from the prior predictive

We go back to the case of braking bikes. Data was $(x_1, y_1), \ldots, (x_5, y_5)$ where x_i was speed and y_i was braking distance.

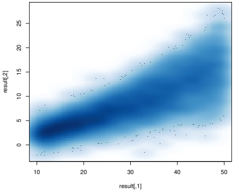
- We now use the model $y_i \mid x_i, a, b, d \sim \text{Normal}(ax_i + bx_i^2, d^2)$ and we need a prior for the three parameters a, b, d.
- For simplicity we try out

  $$a \sim \text{Uniform}[A_0, A_1] \qquad b \sim \text{Uniform}[B_0, B_1] \qquad d \sim \text{Uniform}[D_0, D_1]$$

  for different values $A_0, A_1, B_0, B_1, D_0, D_1$, and simulate from the prior predictive to see if we get something reasonable.
- Values (0.1, 0.3, 0, 0.005, 0.5, 2) produce



or, plotted differently,



Predictions using simulation in a different way

- Sometimes we want to compute probabilities $\pi(y)$ for specific values of y. We can use

$$\begin{aligned}
\pi(y) &= \int \pi(y \mid \theta)\pi(\theta)\, d\theta = \mathsf{E}_\theta\left[\pi(y \mid \theta)\right] \\
\pi(y \mid y_{\text{data}}) &= \int \pi(y \mid \theta)\pi(\theta \mid y_{\text{data}})\, d\theta = \mathsf{E}_{\theta \mid y_{\text{data}}}\left[\pi(y \mid \theta)\right]
\end{aligned}$$

- Idea: Approximate the expectation by generating a sample $\theta_1, \ldots, \theta_N$ from the relevant distribution and average over this sample.
- NOTE: This way to approximate the integral does not suffer from the curse of dimensionality!

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

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

- More generally (assuming the expectation exists)

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

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

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

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

Toy example: The Binomial

We want to predict the probability of 2 successes in 7 trials, with probability of success θ, when $\theta \sim \text{Beta}(7.3, 11.9)$.

- For example, $\text{Beta}(7.3, 11.9)$ could be the posterior after having observed some earlier data.
- Using conjugacy, we can compute

$$\text{Beta-Binomial}(2; 7, 7.3, 11.9) = \binom{7}{2} \frac{B(2 + 7.3, 5 + 11.9)}{B(7.3, 11.9)} = 0.2490633$$

- Using simulation ($N = 10000$) we get (for example) 0.254
- Using Monte Carlo integration ($N = 10000$) we get (for example) 0.2504272

Small example: properties of the posterior

If $\theta = (\alpha, \beta, \gamma)$ is the parameter vector, how do you find the posterior probability that $\alpha > \beta^2$ using Monte Carlo integration?

- We generate a set of vectors $\theta_1, \ldots, \theta_N$ from the posterior for θ given y_{data}.
- Approximate

$$\Pr\left(\alpha > \beta^2 \mid y_{data}\right) \approx \frac{1}{N} \sum_{i=1}^{N} I(\alpha_i > \beta_i^2)$$

where $\theta_i = (\alpha_i, \beta_i, \gamma_i)$.

Example: Approximating quantiles

- Recall: A 95% credibility interval for a random variable θ is an interval so that the probability that θ is in the interval is 95%.
- A possible credibility interval for θ will be $[z_0, z_1]$ where

$$\Pr[\theta < z_0] = 0.025 \qquad \text{and} \qquad \Pr[\theta \leq z_1] = 0.975.$$

- Approximate z_0 and z_1 as follows:
  1. Simulate a sample $\theta_1, \theta_2, \ldots, \theta_N$.
  2. Order it by size to find the 2.5th and 97.5th empirical quantiles.
- In R, use `quantile(theta, c(0.025, 0.975))`.

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

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

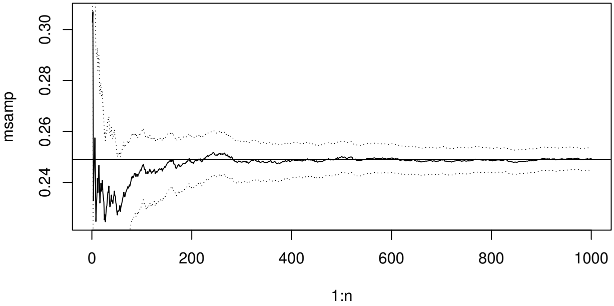
- If we write $\overline{f(\theta)} = \sum_{i=1}^{N} f(\theta_i)/N$ we may approximate

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

Example: Returning to Binomial example

We want to predict the probability of 2 successes in 7 trials, with probability of success θ, when $\theta \sim \text{Beta}(7.3, 11.9)$.

- Find this using Monte carlo integration as follows:
  1. Simulate $\theta_1, \dots, \theta_N$ from Beta(7.3, 11.9).
  2. Compute Binomial$(2; 7, \theta_i)$ for each θ_i
  3. Take the average, and compute the credibility interval as above.
- Showing each result for $N = 1, \dots, 1000$:



Bayesian inference using simulation

- Goal: Compute a probability

$$\pi(y \mid y_{\text{data}}) = \int \pi(y \mid \theta)\pi(\theta \mid y_{\text{data}})\, d\theta = \mathrm{E}_{\theta \mid y_{\text{data}}}\left[\pi(y \mid \theta)\right]$$

- We can do this (also for θ with high dimension!) by
  1. Generating a sample $\theta_1, \ldots, \theta_N \sim \theta \mid y_{\text{data}}$.
  2. Approximating $\pi(y \mid y_{\text{data}}) \approx \sum_{i=1}^{N} \pi(y \mid \theta_i)/N$.
- To solve first step: Find a simulation method for densities known only up to a factor, as

$$\pi(\theta \mid y_{\text{data}}) \propto_\theta \pi(y_{\text{data}} \mid \theta)\pi(\theta).$$

- Today, we continue with more basics on simulation.

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$[0,1]$.

The inverse transform

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

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

- Example: Discrete distributions.
- 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

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

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 $\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$.

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 $\text{Normal}(0, 1/\tau)$.

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 $x_1, \dots, x_k$ are i.i.d Normal$(0, 1)$ then $x = (x_1, \dots, x_n)^t \sim \text{Normal}_k(0, I)$.
- 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.

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.

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.

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
$$f(h(y))|h'(y)|$$
- So: 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.

More about priors

- Alternative 1: Informative prior based on earlier data. (Easy).
- Alternative 2: Informative prior based on "contextual knowledge":
  - Simulate from the prior predictive and assess the result.
  - "Prior elicitation": Get probability statements from an expert, and convert to properties of prior.
- Alternative 3: Non-informative priors:
  - Examples: Gamma$(\tau; 0, 0) = 1/\tau$, or Beta$(\theta; 0, 0) = \frac{1}{\theta(1-\theta)}$.
  - Examples: "Flat" priors like Normal$(\mu; 0, \infty)$ or Beta$(1, 1)$.
  - MAKE SURE YOUR POSTERIOR IS PROPER!
- You may *sometimes* use linear combinations of priors of different types.
- Check that "reasonable" changes in your prior result in small changes in your predictions.
- ...but is there a general theory for non-informative priors?

Different parametrizations using flat priors

Assume a model can be expressed using two alternative parameters, θ and ϕ, related with $\theta = f(\phi)$.

- A prior $\pi_\theta(\theta)$ is transformed to the prior

$$\pi_\phi(\phi) = \pi_\theta(f(\phi))|f'(\phi)|$$

- Example: If $\pi_\theta(\theta) \propto_\theta 1$ and $\theta = \log(\phi)$ with $\phi > 0$ then

$$\pi_\phi(\phi) \propto_\phi \pi_\theta(\log(\phi))\frac{1}{\phi} \propto_\phi \frac{1}{\phi}.$$

- In general, a prior that is "flat" using one parametrization is not flat using another.
- Saying that you use a flat prior is always related to the particular parametrization you use!

Jeffreys prior

- Given a likelihood $\pi(y \mid \theta)$ the Fisher information is defined as

$$\mathcal{I}_\theta(\theta) = \int \left(\frac{\partial}{\partial \theta} \log \pi(y \mid \theta) \right)^2 \pi(y \mid \theta) \, dy.$$

- One can show that, if $\theta = f(\phi)$ then

$$\mathcal{I}_\phi(\phi) = \mathcal{I}_\theta(f(\phi)) \left(f'(\phi)\right)^2.$$

- Thus, defining

$$\pi_\theta(\theta) \propto_\theta \sqrt{\mathcal{I}_\theta(\theta)}$$

  gives a way to define a prior invariant of the parametrization!
- This is Jeffreys prior. It can also be defined for multivariate θ.
- Example: For the Binomial likelihood, Jeffreys prior becomes $\text{Beta}(1/2, 1/2)$!