STAT 535: Chapter 6: Simulation and Markov Chain Monte Carlo (MCMC) Techniques

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2022

The Monte Carlo Method

- The **Monte Carlo method** involves studying a distribution (e.g., a posterior) and its characteristics by generating many random observations having that distribution.
- If $\theta^{(1)}, \ldots, \theta^{(S)} \overset{\text{iid}}{\sim} p(\theta|\boldsymbol{y})$, then the empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ approximates the posterior, when S is large.
- By the **law of large numbers**,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \rightarrow E[g(\theta)|\boldsymbol{y}]$$

as $S \rightarrow \infty$.

The Monte Carlo Method

So as $S \to \infty$:

$$\bar{\theta} = \frac{1}{S}\sum_{s=1}^{S}\theta^{(s)} \to \text{ posterior mean}$$

$$\frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \bar{\theta})^2 \to \text{ posterior variance}$$

$$\frac{\#\{\theta^{(s)} \leq c\}}{S} \to P[\theta \leq c|\boldsymbol{y}]$$

$$\text{median}\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \text{ posterior median}$$

(and similarly for **any** posterior quantile).

The Monte Carlo Method

- If the posterior is a "common" distribution, as in many conjugate analyses, we could draw samples from the posterior using R functions.

Example 1: (General Social Survey)

- **Sample 1**: # of children for women age 40+, no bachelor's degree.
- **Sample 2**: # of children for women age 40+, bachelor's degree or higher.
- Assume Poisson(θ_1) and Poisson(θ_2) models for the data.
- We use gamma(2,1) priors for θ_1 and for θ_2.

The Monte Carlo Method

- **Data**: $n_1 = 111$, $\sum_i y_{i1} = 217$
- **Data**: $n_2 = 44$, $\sum_i y_{i2} = 66$
- $\Rightarrow$ Posterior for θ_1 is gamma(219,112).
- $\Rightarrow$ Posterior for θ_2 is gamma(68, 45).
- Find $P[\theta_1 > \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2]$.
- Find posterior distribution of the ratio $\frac{\theta_1}{\theta_2}$.
- See `R` example using Monte Carlo method on course web page.

Grid Approximation

- In many cases the posterior distribution does not have a simple **recognizable** form, and so we cannot sample from it using built-in `R` functions like "`rgamma`"
- We can still approximate it using simulation techniques such as **grid approximation** or **Markov chain Monte Carlo**.
- We first discuss the simpler approach of grid approximation.
- Let's begin with an example where we know the true posterior: The Gamma-Poisson Model.

Grid Approximation with the Gamma-Poisson

- The book gives a simple example of Poisson data with $n = 2$ observations: $Y_1 = 2$ and $Y_2 = 8$. We choose a Gamma$(3, 1)$ prior for our parameter of interest, λ.
- We know how the derive the posterior distribution analytically in this case (Exercise: Confirm that it is a Gamma$(13, 3)$), but suppose we didn't.
- We could simulate a **grid** of values of the parameter over its range of possible values.
- Since our λ follows a gamma distribution here, it could take values between 0 and ∞, but realistically it is nearly certain to take values between 0 and 15 (see plot of Gamma$(3, 1)$ prior in R).
- So we can generate **many** (say, 501) equally-spaced values of λ between 0 and 15.
- We will then plug these values into our prior $f(\lambda)$ and our likelihood $L(\lambda|\mathbf{y})$.

General Steps for Grid Approximation

- If we have a prior $f(\theta)$ and a likelihood $L(\theta|\mathbf{y})$, here are the steps to approximate the posterior:
  1. Generate a grid of values of θ over its range of possible (or realistic) values.
  2. Plug each value from the grid into the prior $f(\theta)$ and the likelihood $L(\theta|y)$.
  3. Multiply $f(\theta) \times L(\theta|y)$ for each θ-value in the grid.
  4. Then normalize these products so that they sum to 1 (this is done by dividing each value by the sum of the products). The result is an approximation of the posterior probabilities for each θ-value in the grid.
  5. Randomly sample from the grid of θ-values, selecting them based on their normalized posterior probabilities.
- Luckily, this can be done easily and quickly in R.

Grid Approximation in R with the Gamma-Poisson

- Recall the example of Poisson data with $n = 2$ observations: $Y_1 = 2$ and $Y_2 = 8$. We choose a $\text{Gamma}(3, 1)$ prior for our parameter of interest, λ.
- We can generate 501 equally-spaced values of λ between 0 and 15.
- We plug these values into our prior $f(\lambda)$ and our likelihood $L(\lambda|\mathbf{y})$ (easy to do in R).
- We find the normalized posterior probabilities and sample the λ values according to these probabilities (again, easy to do in R).
- See the R code and plots to show how close the approximated posterior comes to the true posterior.
- We can use the Monte Carlo methods to get posterior summary statistics (mean, median, variance, etc.).

MCMC Methods

- Grid approximation tends to break down when the prior and/or likelihood are especially complicated or when there are more than one or two parameters of interest.
- In practical problems, **Markov chain Monte Carlo** (MCMC) sampling methods are used.
- A **Markov chain** is an ordered, indexed set of random variables (a stochastic process) in which the value of each quantity depends probabilistically **only** on the previous quantity.

MCMC Methods

- Specifically, if $\{\theta^{[0]}, \theta^{[1]}, \theta^{[2]}, \ldots\}$ is a Markov chain, then it has the **Markovian** property:
- For any set $\mathcal{A}$,

$$P\{\theta^{[t]} \in \mathcal{A} | \theta^{[0]}, \theta^{[1]}, \ldots, \theta^{[t-1]}\} = P\{\theta^{[t]} \in \mathcal{A} | \theta^{[t-1]}\}$$

- So $\theta^{[t]}$ is **conditionally independent** of all earlier values **except** the previous one.
- So the values in a Markov chain are not independent, but are "almost independent."

Gibbs Sampling

- The **Gibbs Sampler** is a MCMC algorithm that approximates the **joint distribution** of k random quantities by sampling from each **full conditional** distribution in turn.
- **Example**: We are interested in the distribution of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. The Gibbs algorithm is:

1. Choose initial values $\boldsymbol{\theta}^{[0]} = (\theta_1^{[0]}, \theta_2^{[0]}, \ldots, \theta_k^{[0]})$.
2. Cycle through each **full** conditional distribution, sampling, for $t = 1, 2, \ldots$

$$\begin{aligned}
\theta_1^{[t]} &\sim p(\theta_1 | \theta_2^{[t-1]}, \ldots, \theta_k^{[t-1]}) \\
\theta_2^{[t]} &\sim p(\theta_2 | \theta_1^{[t]}, \theta_3^{[t-1]}, \ldots, \theta_k^{[t-1]}) \\
&\vdots \\
\theta_k^{[t]} &\sim p(\theta_k | \theta_1^{[t]}, \theta_2^{[t]}, \ldots, \theta_{k-1}^{[t]})
\end{aligned}$$

3. Repeat steps in (2) until convergence.

Gibbs Sampling

- We must be able to sample from each of the full conditional distributions to use the Gibbs Sampler.
- Note that in each step, the **most recent** value of **each** θ_j is conditioned on.
- After many cycles, the sampled values of $(\theta_1, \ldots, \theta_k)$ will approximate random draws from the joint distribution of $(\theta_1, \ldots, \theta_k)$.
- Then we can summarize, say, a posterior distribution of interest as before.

A Simple Gibbs Example

- **Example 2**: Testing the effectiveness of a seasonal flu shot.
- 20 individuals are given a flu shot at the start of winter.
- At the end of winter, follow up to see whether they contracted flu.

  Let

$$X_i = \begin{cases} 1 & \text{if shot effective (no flu)} \\ 0 & \text{if ineffective (contracted flu)} \end{cases}$$

- Suppose the 20th individual was unavailable for followup.
- Define $Y = \sum_{i=1}^{19} X_i$.

A Simple Gibbs Example

- If θ is the probability the shot is effective, then

$$p(y|\theta) = \binom{19}{y} \theta^y (1-\theta)^{19-y}$$

- If we had the complete data (for Y **and** X_{20}), then

$$p(\theta|y, x_{20}) = \binom{20}{y + x_{20}} \theta^{y+x_{20}} (1-\theta)^{20-y-x_{20}}$$

- If we put in "temporary" values θ^* and x_{20}^* for the unknown quantities, then

$$\theta | X_{20}^*, Y \sim \text{beta}(Y + X_{20}^* + 1, 20 - Y - X_{20}^* + 1)$$

$$\text{and } X_{20}|Y, \theta^* \sim \text{Bernoulli}(\theta^*)$$

A Simple Gibbs Example

- We can repeatedly sample from these “full conditional” distributions and eventually get a sample from the joint distribution of (θ, X_{20}).
- See R example with data.

A More Complicated Gibbs Example (Changepoint)

Example 3: (Coal Mining Disasters)

- Data are yearly counts of British coal mine disasters, 1851-1962.
- Relatively large counts in the early era, small counts in the later years.
- **Question**: When did the mean of the process change?
- We model the data using two Poisson distributions:
- "Early" data: $Y_1, \ldots, Y_k | \lambda \overset{\text{iid}}{\sim} \text{Pois}(\lambda), \;\; i = 1, \ldots, k$
- "Later" data: $Y_{k+1}, \ldots, Y_n | \phi \overset{\text{iid}}{\sim} \text{Pois}(\phi), \;\; i = k+1, \ldots, n$
- We must estimate each Poisson mean, λ and ϕ, and **also** the "changepoint" k.

A More Complicated Gibbs Example (Changepoint)

Consider the priors:

$$\begin{aligned} \lambda &\sim \text{gamma}(\alpha, \beta) \\ \phi &\sim \text{gamma}(\gamma, \delta) \\ k &\sim \text{discrete uniform on}\{1, 2, \dots, n\} \end{aligned}$$

- If we believe the mean annual disaster count for early years is ≈ 4 and for later years is ≈ 0.5, let $\alpha = 4$, $\beta = 1$, $\gamma = 1$, $\delta = 2$ be the hyperparameters.

A More Complicated Gibbs Example (Changepoint)

Then the posterior is $p(\lambda, \phi, k|\boldsymbol{y})$

$$\propto L(\lambda, \phi, k|\boldsymbol{y}) p(\lambda) p(\phi) p(k)$$

$$= \left[\prod_{i=1}^{k} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right] \left[\prod_{i=k+1}^{n} \frac{e^{-\phi}\phi^{y_i}}{y_i!}\right] \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\right] \left[\frac{\delta^\gamma}{\Gamma(\gamma)}\phi^{\gamma-1}e^{-\delta\phi}\right] \left[\frac{1}{n}\right]$$

$$\propto e^{-k\lambda}\lambda^{\sum_{i=1}^{k} y_i} e^{-(n-k)\phi}\phi^{\sum_{k+1}^{n} y_i} \lambda^{\alpha-1}e^{-\beta\lambda}\phi^{\gamma-1}e^{-\delta\phi}$$

$$= \lambda^{\alpha+\sum_{i=1}^{k} y_i - 1} e^{-(\beta+k)\lambda} \phi^{\gamma+\sum_{k+1}^{n} y_i - 1} e^{-(\delta+n-k)\phi}$$

So full conditionals are:

$$\lambda|\phi, k \sim \text{gamma}(\alpha + \sum_{i=1}^{k} y_i, \beta + k)$$

$$\phi|\lambda, k \sim \text{gamma}(\gamma + \sum_{i=k+1}^{n} y_i, \delta + n - k)$$

A More Complicated Gibbs Example (Changepoint)

To get the full conditional for k, note the joint density of the data is:

$$
\begin{aligned}
p(\boldsymbol{y}|k,\lambda,\phi) &= \left[\prod_{i=1}^{k} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right]\left[\prod_{i=k+1}^{n} \frac{e^{-\phi}\phi^{y_i}}{y_i!}\right] \\
&= \left[\prod_{i=1}^{n} \frac{1}{y_i!}\right] e^{k(\phi-\lambda)} e^{-n\phi} \lambda^{\sum\limits_{i=1}^{k} y_i} \left[\prod_{i=k+1}^{n} \phi^{y_i}\right]\left[\frac{\prod\limits_{i=1}^{k} \phi^{y_i}}{\phi^{\sum_{i=1}^{k} y_i}}\right] \\
&= \left[\prod_{i=1}^{n} \frac{e^{-\phi}\phi^{y_i}}{y_i!}\right]\left[e^{k(\phi-\lambda)}\left(\frac{\lambda}{\phi}\right)^{\sum\limits_{i=1}^{k} y_i}\right] \\
&= f(\boldsymbol{y},\phi)g(\boldsymbol{y}|k)
\end{aligned}
$$

A More Complicated Gibbs Example (Changepoint)

By Bayes' Law, for any particular value k^* of k,

$$p(k^*|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, \phi)g(\boldsymbol{y}|k^*)p(k^*)}{\sum_{k=1}^{n} f(\boldsymbol{y}, \phi)g(\boldsymbol{y}|k)p(k)}$$

Since $p(k) = 1/n$ (constant), we have

$$p(k^*|\boldsymbol{y}) = p(k^*|\boldsymbol{y}, \lambda, \phi) \propto \frac{g(\boldsymbol{y}|k^*)}{\sum_{k=1}^{n} g(\boldsymbol{y}|k)}$$

(full conditional for k)

- This ratio defines a probability vector for k that we use at each iteration to sample a value of k from $\{1, 2, \dots, n\}$.
- see R example (Coal mining data)

Metropolis-Hastings Sampling

- When the full conditionals for each parameter cannot be obtained easily, another option for sampling from the posterior is the Metropolis-Hastings (M-H) algorithm.
- The M-H algorithm also produces a **Markov chain** whose values approximate a sample from the posterior distribution.
- For this algorithm, we need the form (except for a normalizing constant) of the posterior $p(\cdot|\boldsymbol{y})$ for $\boldsymbol{\theta}$, the parameter(s) of interest.
- We also need a **proposal** (or **instrumental**) distribution $q(\cdot|\cdot)$ that is easy to sample from.

Metropolis-Hastings Sampling

- The M-H algorithm first specifies an initial value for $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{[0]}$. Then:
- After iteration t, suppose the most recently drawn value is $\boldsymbol{\theta}^{[t]}$.
- Then sample a candidate value $\boldsymbol{\theta}^*$ from the proposal density.
- Let the $(t+1)$-st value in the chain be

$$\boldsymbol{\theta}^{[t+1]} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min\{a(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{[t]}), 1\} \\ \boldsymbol{\theta}^{[t]} & \text{with probability } 1 - \min\{a(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{[t]}), 1\} \end{cases}$$

where

$$a(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{[t]}) = \frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})}{p(\boldsymbol{\theta}^{[t]}|\boldsymbol{y})} \frac{q(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t]})}$$

is the "acceptance ratio."

Metropolis-Hastings Sampling

- In practice we would accomplish this by sampling $U^{[t]} \sim U(0,1)$ and choosing $\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^*$ if $a(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{[t]}) > u^{[t]}$; otherwise choose $\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]}$.
- Note that if the proposal density $q(\cdot|\cdot)$ is **symmetric** such that $q(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t]})$, then the acceptance ratio is simply

$$\frac{p(\boldsymbol{\theta}^*|\mathbf{y})}{p(\boldsymbol{\theta}^{[t]}|\mathbf{y})}.$$

Metropolis-Hastings Example

Example 5 (Sparrow data): We gather data on a sample of 52 sparrows:

$$X_i = \text{age of sparrow (to nearest year)}$$
$$Y_i = \text{Number of offspring that season}$$

- We expect that the offspring number rises and then falls with age, so we assume a quadratic trend.
- We model the number of offspring at a given age x as Poisson:

$$Y|x \sim \text{Pois}(\mu_x)$$

Metropolis-Hastings Example

- Since we know μ_x must be positive, we use the model:

$$E[Y|x] = e^{\beta_0 + \beta_1 x + \beta_2 x^2}$$

- This Poisson regression model is a **generalized linear model** (GLM).
- Our parameter of interest is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$.
- But note that conjugate priors do not exist for **non-normal** GLMs.
- We will use the M-H algorithm to sample from our posterior.

Metropolis-Hastings Example

- Let the prior on β be multivariate normal with **independent** components:

$$\beta \sim MVN(\mathbf{0}, \mathbf{\Sigma}), \text{ where } \mathbf{\Sigma} = 100 \times \boldsymbol{I}_3$$

- We will choose our **proposal** density to be multivariate normal with mean vector $\beta^{[t]}$ (the current value).
- The covariance matrix of the proposal density is sort of a tuning parameter. We will choose

$$\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \text{ where } \hat{\sigma}^2 = \text{var}\{\ln(y_1 + 0.5), \ldots, \ln(y_n + 0.5)\}.$$

- We can adjust this if our **acceptance rate** is too high or too low.
- Usually we like an acceptance rate between 20% and 50%.

Metropolis-Hastings Example

- Since our proposal density is symmetric, our acceptance ratio is simply

$$\frac{p(\beta^*|\boldsymbol{X}, \boldsymbol{y})}{p(\beta^{[t]}|\boldsymbol{X}, \boldsymbol{y})} = \frac{L(\beta^*|\boldsymbol{X}, \boldsymbol{y})p(\beta^*)}{L(\beta^{[t]}|\boldsymbol{X}, \boldsymbol{y})p(\beta^{[t]})}$$

$$= \frac{\prod_{i=1}^{n} \text{dpois}(y_i, \exp[\boldsymbol{x}_i^T \beta^*]) \prod_{j=1}^{3} \text{dnorm}(\beta_j^*, 0, 10)}{\prod_{i=1}^{n} \text{dpois}(y_i, \exp[\boldsymbol{x}_i^T \beta^{[t]}]) \prod_{j=1}^{3} \text{dnorm}(\beta_j^{[t]}, 0, 10)}$$

  where the Poisson density `dpois` and the normal density `dnorm` can be found easily in `R`.

- See `R` example with real sparrow data.

Other Metropolis-Hastings Issues

- In practice, it is recommended to check the acceptance rate (the proportion of proposed $\boldsymbol{\beta}^*$ values that are "accepted").
- We also check the serial correlation of the $\left\{\beta_j^{[t]}\right\}$ values using a plot of the **autocorrelation function**.
- If the values do not "appear" independent, we can alleviate this by choosing every k th value in the chain as our posterior sample (**thinning**).
- A **trace plot** is a plot of the sampled parameter values over the iterations of the algorithm. We use this to assess whether the algorithm has "converged" and can be assumed to be sampling from the actual posterior distribution.
- Ideally, we'd like to see a trace plot that looks like a "hairy caterpillar" after a sufficient number of iterations.