Basic of Markov chain simulation

Jongjin, Lee

October 13, 2020

Seoul National University

Table of Contents

Table of Contents

Posterior quantity.

- Posterior quantity.

$$E[g(\theta)|y] = \int_\theta g(\theta)p(\theta|y)d\theta$$

  - Posterior predictive distribution.
  - Model checking. (Ch7)
- If the calculation of the posterior distribution is infeasible, how to calculate posterior quantity.

Monte Carlo Method

- Posterior quantity can be approximated by sampling from posterior distribution.

$$E[g(\theta)|y] = \int_\theta g(\theta)p(\theta|y)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)})$$

- Then, how to draw independent samples from posterior distribution?

Monte Carlo Methods

- Markov chain simulation (Markov chain Monte Carlo, MCMC)
- Gibbs sampling / Metropolis algorithm / Metropolis-Hastings algorithm.
  - The sampling is done sequentially, with the distribution of the sampled draws depending on the last value drawn.

Table of Contents

Markov chain

- A sequence of random variables $X^0, X^1, \dots$ is a Markov chain if
$$p(X^t|X^0, \dots, X^{t-1}) = p(X^t|X^{t-1})$$

- $p(X^t|X^{t-1})$ is called as a transition probability(transition kernel).
- If $p(X^t|X^{t-1})$ does not depend on t, then Markov chain is called homogeneous.
- For a homogeneous Markov chain, we will denote transition probability as $p(y|x) = P(X^1 = y|X^0 = x)$

Stationary distribution

- For a homogeneous Markov chain with $p(y|x)$, a distribution $\pi(y)$ which satisfies

$$\pi(y) = \int p(y|x)\pi(x)dx$$

is called a stationary distribution.

- A stationary distribution may not exist or may not be unique.

Example

- Suppose the space are (Rain, Sunny, Cloudy) and weather follows a Markov process
- The transition probability

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

- Suppose that today is sunny, $\pi(0) = (0, 1, 0)$, what is the expected weather two days later, or seven days?

$$\pi(2) = \pi(0)\mathbf{P}^2 = (0.375, 0.25, 0.375)$$
$$\pi(7) = \pi(0)\mathbf{P}^7 = (0.400024, 0.199951, 0.400024)$$

Example

- After a sufficient amount of time, the expected weather becomes independent of the initial value.
- The chain has reached a stationary distribution.
- Stationary distribution π^*

$$\pi^* = \pi^* \mathbf{P} = (0.4, 0.2, 0.4)$$

- We will consruct a Markov chain which has target distribution (posterior distribution) as a stationary distribution.

Ergodic Markov chains

- Ergodicity Theorem
  - If a Markov chain is ergodic, then a unique stationary distribution π^* exists, which is independent of the initial state.
- Ergodic Markov chain
  - irreducible/recurrent nonnull(positive)/aperiodic
  - aperiodic and recurrent nonnull $\rightarrow$ existence of π.
  - irreducible $\rightarrow$ uniqueness of π.

MCMC(Markov chain Monte Carlo)

- The Monte Carlo Method.

$$E[g(\theta)|y] = \int_\theta g(\theta)p(\theta|y)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)})$$

  - Independent samples
- MCMC
  - Construct irreducible, aperiodic, positive Markov chain with stationary distribution $p(\theta|y)$.
  - Simulate $\theta^{(1)}, \theta^{(2)}, \ldots$ from markov chain. Then:

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to E[g(\theta)|y] \text{ as } S \to \infty$$

Table of Contents

Markov chain simulations

- Gibbs sampler / Metropolis algorithm / Metropolis-Hasting algorithm
- We denote samples at each iteration as $\theta^t, t = 0, 1, \ldots,$
- For each t, θ^t is sampled from a certain transition distribution $T_t(\theta^t|\theta^{t-1})$
- The transition probability distributions must be constructed so that Markov chain converges to a unique stationary distribution, $p(\theta|y)$.
- A variety of Markov chain can be constructed.

Gibbs sampler

- Useful in many multidimensional problems
- Alternating conditional sampling
- Generating samples from joint distribution is difficult, but generating samples from condition distribution is easy.

Gibbs sampler

- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$
- Each θ_j could be a subvector of $\boldsymbol{\theta}$(dim ≥ 1)
- $\boldsymbol{\theta}_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_d)$
- Drawing each subset of $\boldsymbol{\theta}$ conditional on the value of all the others.
- In each iteration t,
    - $p(\theta_j | \boldsymbol{\theta}_{-j}^{t-1}, y)$

Gibbs sampler

- Gibbs sampler
- For $t = 1$ to S
  1. Generate $\theta_1^t \sim p(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)}, y)$
  2. Generate $\theta_2^t \sim p(\theta_2|\theta_1^{(t)}, \theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)}, y)$
  3. $\cdots$
  4. Generate $\theta_d^t \sim p(\theta_2|\theta_1^{(t)}, \ldots, \theta_2^{(t)}, y)$
- This Markov chain has posterior distribution as a stationary distribution.

Example: Bivariate normal distribution

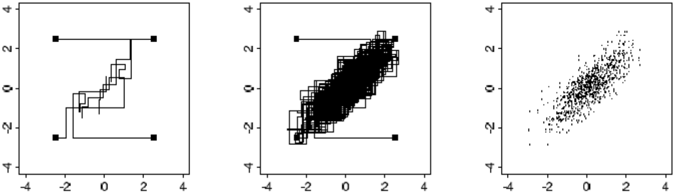
- Posterior distribution.

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \mid y \sim \mathrm{N}\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

- Conditional distribution

$$\theta_1 \mid \theta_2, y \sim \mathrm{N}\left(y_1 + \rho\left(\theta_2 - y_2\right), 1 - \rho^2\right)$$
$$\theta_2 \mid \theta_1, y \sim \mathrm{N}\left(y_2 + \rho\left(\theta_1 - y_1\right), 1 - \rho^2\right)$$

Example: Bivariate normal distribution



- $\rho = 0.8, (y_1, y_2) = (0, 0)$, four independent sequences started at $(\pm 2.5, \pm 2.5)$

Metropolis algorithm

- Draw values of θ from approximate distributions and correctthose draws to better approximate the target distribution.
- Random walk with an acceptance/rejection rule.
- Symmetric jumping distribution(proposal distribution)
- $T_t(\theta^t|\theta^{t-1})$ is a weighted version of $J_t(\theta^t|\theta^{t-1})$

Metropolis algorithm

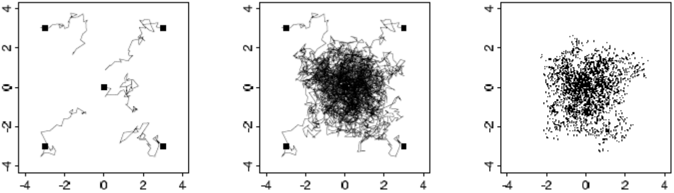
1. Draw a staring point θ^0, ($p(\theta^0|y) > 0$), from starting distribution $p_0(\theta)$
2. For $t = 1$ to S
   1. Sample a proposal θ^* from a jumping distribution(proposal distribution) at time t, $J(\theta^*|\theta^{t-1})$.(symmetric)
   2. Calculate the ratio of the densities
   $$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$$
   3. Set
   $$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

Example: Bivariate unit normal density with normal jumping kernel

- $p(\theta \mid y) = \mathrm{N}(\theta \mid 0, I)$, where I is the 2×2 identity matrix.
- $J_t\left(\theta^* \mid \theta^{t-1}\right) = \mathrm{N}\left(\theta^* \mid \theta^{t-1}, 0.2^2 I\right)$
- $r = \mathrm{N}\left(\theta^* \mid 0, I\right) / \mathrm{N}\left(\theta^{t-1} \mid 0, I\right)$
- In Ch12, we discuss how to set the jumping scale to optimize the efficiency of the Metropolis algorithm.

The Markov simulation.

- Five simulation runs starting from different points.



The sketch of the proof of the validity of the Metropolis algorithm

- Why does the Metropolis algorithm work?
- First, it is shown that the simulated sequence is a Markov chain with a unique stationary distribution.
- Second, The stationary distribution equals this target distribution.

The sketch of the proof of the validity of the Metropolis algorithm

- Ergodicity is from random work.
- Need to show that the posterior distribution is the stationary distribution of this Markov chain
- Consider starting the algorithm at time $t-1$
- Any two such points θ_a, θ_b, drawn from $p(\theta|y)$ and labeled so that $p(\theta_b|y) \geq p(\theta_a|y)$.

The sketch of the proof of the validity of the Metropolis algorithm

- Unconditional probabilty density of a transition form θ_a to θ_b is

$$p(\theta^{t-1} = \theta_a, \theta^t = \theta_b) = p(\theta_a|y)J_t(\theta_b|\theta_a)$$

- Unconditional probabilty density of a transition form θ_b to θ_a is

$$\begin{aligned} p\left(\theta^t = \theta_a, \theta^{t-1} = \theta_b\right) &= p\left(\theta_b \mid y\right) J_t\left(\theta_a \mid \theta_b\right) \left(\frac{p\left(\theta_a \mid y\right)}{p\left(\theta_b \mid y\right)}\right) \\ &= p\left(\theta_a \mid y\right) J_t\left(\theta_b \mid \theta_a\right) \end{aligned}$$

- Since their joint distribution is symmetric, θ^t and θ^{t-1} have the same marginal distributions, and so $p(\theta|y)$ is the stationary distribution of the Markov chain of θ

The Metropolis-Hasting algorithms

- The Metropolis algorithm is a special case of the Metropolis-Hasting algorithm.
- Asymmetric jumping distribution.
- $r = \frac{p(\theta^*|y)/J_t\left(\theta^*|\theta^{t-1}\right)}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}$

Using Gibbs and Metropolis as building blocks

- Gibbs: conditionally conjugate model
- Metropolis: not conditionally conjugate model.
- A general problem with conditional sampling algorithms is that they can be slow when parameters are highly correlated in the target distribution. (reparametriztion or more advanced algorithms)

A good jumping distribution.

- For any θ, it is easy to sample from $p(\theta^*|\theta)$
- Easy to compute the ratio r
- Each jump goes a reasonable distance in the parameter space(otherwise the random walk moves too slowly.)
- The jumps are not rejected too frequently.

Gibbs sampler & Metropolis-Hastings algorithm

- Gibbs sampler can be viewed as special case of the Metropolis-Hastings algorithms
- Define iteration t to consist of a series of d steps.

Gibbs sampler & Metropolis-Hastings algorithm

- Jumping distribution $J_{j,t}(\cdot|\cdot)$. ate step j of iteration t.

$$J_{j,t}^{\text{Gibbs}}\left(\theta^{*} \mid \theta^{t-1}\right)=\begin{cases} p\left(\theta_{j}^{*} \mid \theta_{-j}^{t-1}, y\right) & \text{if } \theta_{-j}^{*}=\theta_{-j}^{t-1} \\ 0 & \text{otherwise} \end{cases}$$

- The ratio at jth step of iteration t is

$$\begin{aligned} r &= \frac{p\left(\theta^{*} \mid y\right) / J_{j,t}^{\text{Gibbs}}\left(\theta^{*} \mid \theta^{t-1}\right)}{p\left(\theta^{t-1} \mid y\right) / J_{j,t}^{\text{Gibbs}}\left(\theta^{t-1} \mid \theta^{*}\right)} \\ &= \frac{p\left(\theta^{*} \mid y\right) / p\left(\theta_{j}^{*} \mid \theta_{-j}^{t-1}, y\right)}{p\left(\theta^{t-1} \mid y\right) / p\left(\theta_{j}^{t-1} \mid \theta_{-j}^{t-1}, y\right)} \\ &= \frac{p\left(\theta_{-j}^{t-1} \mid y\right)}{p\left(\theta_{-j}^{t-1} \mid y\right)} \\ &\equiv 1 \end{aligned}$$

Table of Contents

The basic method of inference

- Use the collection of all the simulated draws from $p(\theta|y)$ to summarize the posterior quantity.
- Two challenges
  - grossly unrepresentative of the target distribution.
  - within sequence correlation.

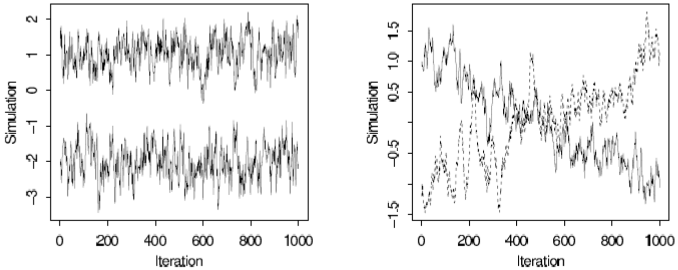
The basic method of inference

- Discarding early iterations(warm-up/burn-in) of simulations.
- Once approximate convergence has been reached, keeping evert kth simulation draw from each sequence and discarding the rest.(Thinning)
- Then, how to assess that the convergence has been reached?

Assessing convergence

- Multiple chains with starting points dispersed throughout parameter space.
  - Stationary and mixing.
  - Within- and between- variance of scalar estimands(posterior quantity).

- Two challenges of monitoring convergence of iterative simulations.
- Stationary and mixing.



Monitoring scalar estimands

- All parameters in the model and any other quantities of interest.
- it is often useful to monitor the value of the logarithm of the posterior density.

Assessing mixing using between and within sequence variances

- We denote interested scalar estimands as ψ
- For calculating between and within sequence variance, discard the first half of each simulation chain as warm-up
- Split each into two same length of sequence.
- *m*: The twice number of chains
- *n*: The length of remained chain each chains
- Suppose we simulate 5 chains, each of length 1000, and then m = 10, n = 250

Assessing mixing using between and within sequence variances

- Between- and within sequence variances

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\psi}_{.j} - \bar{\psi}_{..}\right)^2, W = \frac{1}{m} \sum_{j=1}^{m} s_j^2$$

  where $\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}, \quad \bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$, and $s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\psi_{ij} - \bar{\psi}_{.j}\right)^2$
- We can estimate $\mathrm{var}(\psi \mid y)$

$$\widehat{\mathrm{var}}^{+}(\psi \mid y) = \frac{n-1}{n} W + \frac{1}{n} B$$

Assessing mixing using between and within sequence variances

- $\widehat{\mathrm{var}}^{+}(\psi \mid y)$ overestimates the marginal posterior variance assuming the starting distribution is appropriately overdispersed.
- W is an underestimate of $\mathrm{var}(\psi \mid y)$
- W approaches $\mathrm{var}(\psi \mid y)$ as $n \to \infty$
- Potential scale reduction.

$$\widehat{R} = \sqrt{\frac{\widehat{\mathsf{var}}(\psi \mid y)}{W}}$$

  which declines to 1 as $n \to \infty$
- If $\hat{R}$ is high, further simulations may improve our inference.

Table of Contents

MCMC samples are dependent

- MCMC samples are dependent
- This does not effect the validity of inference on the posterior, if samplers has time to explore the posterior distributions.
- Highly correlated MCMC samplers requires more samples to produce the same level of Monte Carlo for an estimate

Effective number of simulation draws

- Effective sample size is some sort of "exchange rate" between dependent and independent samples.
- The number of effectively indepedent draws from the posterior distribution that the Markov chain is equivalent to.
- The larger the better.
- They suggest running the simulation until n_{eff} is at least m

Effective number of simulation draws

- It is usual to compute effective sample size using the following asymptotic formula for the variance of the average of a correlated sequence.

$$\lim_{n\to\infty} mn \,\mathrm{var}\left(\bar{\psi}_{..}\right) = \left(1 + 2\sum_{t=1}^{\infty} \rho_t\right) \mathrm{var}(\psi \mid y)$$

  ρ_t is the autocorrelation of the sequence ψ at lag t.

- If the simulation draws were independent, the effective sample size is mn

$$\mathrm{var}\left(\bar{\psi}_{..}\right) = \frac{1}{mn}\mathrm{var}(\psi \mid y)$$

- Then, in the presence of correlation the effective sample size is

$$n_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{\infty} \rho_t}$$

Effective number of simulation draws

- Compute the total variance using the $\widehat{\text{var}}^+(\psi \mid y)$
- Estimate the correlations by first computing the variogram V_t at each lag t

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\psi_{i,j} - \psi_{i-t,j})^2$$

- We then estimate the correlations by inverting the formula, $\mathrm{E}\left(\psi_i - \psi_{i-t}\right)^2 = 2\left(1 - \rho_t\right)\text{var}(\psi)$

$$\widehat{\rho}_t = 1 - \frac{V_t}{2\widehat{\mathrm{var}}^+}$$

- We compute a partial sum, starting from lag 0 and continuing the sum of autocorrelation estimates for two succesive lags $\hat{\rho}_{2t'} + \hat{\rho}_{2t'+1}$ is negative.

$$\hat{n}_{\mathrm{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t}$$

Effective number of simulation draws

- This convergence diagnostic are based on means and variances, therefore it is vulnerable to the posterior distribution is far from Gaussian
- Using transformations before computing the potential scale reduction factor $\hat{R}$ and the effective sample size n_{eff}

Table of Contents

Example: hierarchical normal model

- Likelihood (y_{ij})

$$\prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}\left(y_{ij} \mid \theta_j, \sigma^2\right)$$

- Prior
  - θ_j from normal distribution with unknown mean μ and variance τ^2
  - $(\mu, \log \sigma, \log \tau) \propto \tau$
- Posterior

$$p(\theta, \mu, \log \sigma, \log \tau \mid y) \propto \tau \prod_{j=1}^{J} \mathrm{N}\left(\theta_j \mid \mu, \tau^2\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}\left(y_{ij} \mid \theta_j, \sigma^2\right)$$

Example: hierarchical normal model

- Initialize (ch13.)
- Gibbs sampler
  - The conditional distribution of each θ_j, normal
  - The conditional distribution of μ, normal
  - The conditional distribution of σ^2, inverse gamma
  - The conditional distribution of τ^2, inverse gamma

Example: hierarchical normal model

- Posterior

$$p(\theta, \mu, \log \sigma, \log \tau \mid y) \propto \tau \prod_{j=1}^{J} \mathrm{N}\left(\theta_j \mid \mu, \tau^2\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}\left(y_{ij} \mid \theta_j, \sigma^2\right)$$

- The conditional distribution of eachθ_j

$$\theta_j \mid \mu, \sigma, \tau, y \sim \mathrm{N}\left(\hat{\theta}_j, V_{\theta_j}\right)$$

where,

$$\hat{\theta}_j = \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_{.j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

$$V_{\theta_j} = \frac{1}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

Example: hierarchical normal model

- Posterior

$$p(\theta, \mu, \log\sigma, \log\tau \mid y) \propto \tau \prod_{j=1}^{J} \mathrm{N}\left(\theta_j \mid \mu, \tau^2\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}\left(y_{ij} \mid \theta_j, \sigma^2\right)$$

- The conditional distribution of μ

$$\mu \mid \theta, \sigma, \tau, y \sim \mathrm{N}\left(\hat{\mu}, \tau^2/J\right)$$

where,

$$\hat{\mu} = \frac{1}{J}\sum_{j=1}^{J} \theta_j$$

Example: hierarchical normal model

- Posterior

$$p(\theta, \mu, \log\sigma, \log\tau \mid y) \propto \tau \prod_{j=1}^{J} \mathrm{N}\left(\theta_j \mid \mu, \tau^2\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}\left(y_{ij} \mid \theta_j, \sigma^2\right)$$

- The conditional distribution of σ^2

$$\sigma^2 \mid \theta, \mu, \tau, y \sim \mathrm{Inv}-\chi^2\left(n, \hat{\sigma}^2\right)$$

where,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(y_{ij} - \theta_j\right)^2$$

Example: hierarchical normal model

- Posterior

$$p(\theta, \mu, \log\sigma, \log\tau \mid y) \propto \tau \prod_{j=1}^{J} \mathrm{N}\left(\theta_j \mid \mu, \tau^2\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}\left(y_{ij} \mid \theta_j, \sigma^2\right)$$

- The conditional distribution of τ^2

$$\tau^2 \mid \theta, \mu, \sigma, y \sim \text{Inv}-\chi^2\left(J-1, \hat{\tau}^2\right)$$

where,

$$\hat{\tau}^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j - \mu)^2$$

References i