Markov Chain Monte Carlo

How to Obtain Posterior

- Posterior distribution is proportional to a product of two distributions (prior and likelihood)
  - Conjugate – we know the class of the posterior distribution
  - Not conjugate – we don't know the distribution, simulation methods have to be adopted, Markov Chain Monte Carlo

Markov Chain

- Markov Chain
  - Start with an initial point θ_0
  - The next point only relies on the value of the current point, using conditional probability $p(\theta_{r+1}|\theta_r)$, not the earlier stages
  - A nice property of Markov Chain is that, under certain conditions, the distribution of $(\theta_r|\theta_0)$ will converge to a stationary/invariant/equilibrium distribution.
    - This distribution will have no influence by the value of θ_0.

Markov Chain Monte Carlo

- Markov Chain Monte Carlo
  - We want to establish the transition matrix $p(\theta_{r+1}|\theta_r)$, so that the simulated chain has a stationary distribution that is the posterior distribution we want.
  - The chain is ***Ergodic*** (遍历性), if it satisfies:

$$\lim_{R\to\infty} \frac{1}{R} \sum h(\theta_r) = E[h(\theta)]$$

- Law of large numbers
  - Standard version requires independence of the draws
  - As long as the dependence is not too strong, it still works

Example of Markov Chain

Consider the Discrete case

$$S = \left\{ q^1, q^2, \ldots, q^d \right\}$$

d possible states – such as a discretization of parameter space

define a sequence of r.v.s given some initial starting point.

$$\Pr\left[\theta_{r+1} = \theta^j \middle| \theta_r = \theta^i \right] = p_{i,j}$$

prob chain is in state j at "time" or iteration r+1 given in state i at iteration r

$$P = \left[p_{i,j} \right]; \ \sum_j p_{i,j} = 1$$

the rows of P specify the conditional distribution of θr+1 given realized value of θr

5

Stationary Distribution

Suppose we specify a distribution for θ_0 , π_0. What is the distribution of θ_1?

$$\Pr\left[q_1 = q^j \right] = \sum_i \Pr\left[q_0 = q^i \right] \Pr\left[q_1 = q^j \middle| q_0 = q^i \right]$$

$$= \sum_i p_{0,i} p_{i,j}$$

or

$$p_1 = p_0 P, \ p_2 = p_1 P = p_0 P^2, \ldots$$

row vectors

As r increases, the effects of starting distribution should "wear" off.

Stationary Distribution

If we start in the stationary distribution, we should stay there which is why the stationary distribution is sometimes called the invariant distribution

$$\pi P = \pi$$

eigenvector corresponding to eigenvalue of 1.

$$\text{Let } P = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} \qquad \pi = [1/3 \quad 2/3]$$

$$\pi P = [1/3 \quad 2/3]\begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = [1/3 \quad 2/3]$$

Stationary Distribution

If pi,j > 0, then chain is called irreducible and there will be a stationary distribution

$$\lim_{r\to\infty} \pi_0 P^r = \pi$$

An irreducible chain is free to navigate. That is, we can get from anywhere to anywhere else (not necessarily in one step, though).

Ex:

$$P = \begin{bmatrix} .5 & .1 & .4 \\ 0 & .5 & .5 \\ 0 & .4 & .6 \end{bmatrix}$$

"trapped" in states 2 and 3

Practical Considerations

- Effect of initial conditions
  - “burn-in” – run for B iterations, discard and use only the las R-B
- Non-iid Simulate – is this a problem?
  - No, Law of Large Numbers works for dependent sequences
  - Yes, simulation error larger than IID sequence
- How to construct the chain?

First Method: Gibbs Sampler

- Posterior distribution for model parameters could be high-dimensional, which is hard to be simulated from
- Gibbs sampler provides a method to draw from the conditionals in order to achieve the joint.
- An example: to simulate the joint distribution of (θ_1, θ_2)
  - Start with (θ_1^0, θ_2^0)
  - Draw the next θ_1^{r+1} from the conditional distribution $\theta_1|\theta_2^r$
  - Draw the next θ_2^{r+1} from the conditional distribution $\theta_2|\theta_1^{r+1}$

Why does it work?

- Suppose $(\theta_1^r, \theta_2^r) \sim \pi(\cdot)$, we need to prove that $(\theta_1^{r+1}, \theta_2^{r+1}) \sim \pi(\cdot)$
- $(\theta_1^r, \theta_2^r) \sim \pi(\cdot)$, therefore $\theta_1^r \sim \pi_1(\cdot)$, marginal distribution
- $\theta_2^{r+1} \sim \pi_{2|1}(\theta_2 | \theta_1^r)$ in Gibbs sampler
- Combine the last two steps, we get $\theta_2^{r+1} \sim \int \pi_{2|1}(\theta_2|\theta_1)\pi_1(\theta_1)d\theta_1 = \pi_2(\quad)$
- Same to prove $\theta_1^{r+1} \sim \pi_1(\quad)$
- Therefore $(\theta_1^{r+1}, \theta_2^{r+1}) \sim \pi(\cdot)$

Gibbs Sampler

$$\theta \sim N\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

$$\theta_1 | (\theta_2 = a) \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(a - \mu_2), \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right)$$

$$q_2 | q_1 \sim N\left(r\, q_1, \left(1 - r^2\right)\right) \qquad q_1 | q_2 \sim N\left(r\, q_2, \left(1 - r^2\right)\right)$$

A simulator: Start at point $\theta^0 = \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix}$

Draw θ^1 in two steps:

$$q_2^1 \sim N\left(r\, q_1^0, 1 - r^2\right)$$

$$q_1^1 \sim N\left(r\, q_2^1, 1 - r^2\right)$$

Note: this is a Markov Chain. Current point entirely summarizes past.

Gibbs Sampler

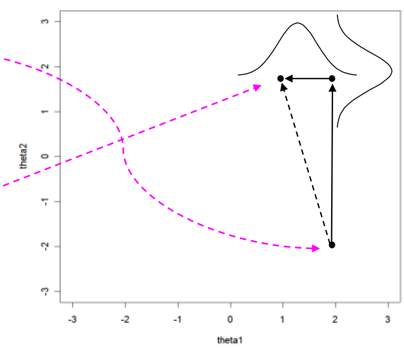
A simulator:

Start at point $\theta^0 = \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix}$

Draw θ^1 in two steps:

$$\theta_2^1 \sim N\left(\rho\theta_1^0, 1-\rho^2\right)$$

$$\theta_1^1 \sim N\left(\rho\theta_2^1, 1-\rho^2\right)$$



repeat!

General Gibbs Sampler

$\theta = (\theta_1, \theta_2, \dots, \theta_p)$, "blocking"

Sample from

$$\theta_{t+1,1} \sim f_1\left(\theta_1 | \theta_{t,2}, \theta_{t,3}, \dots, \theta_{t,p}\right)$$

$$\theta_{t+1,2} \sim f_2\left(\theta_2 | \theta_{t+1,1}, \theta_{t,3}, \dots, \theta_{t,p}\right)$$

$$\theta_{t+1,3} \sim f_2\left(\theta_3 | \theta_{t+1,1}, \theta_{t+1,2}, \dots, \theta_{t,p}\right)$$

$$\dots$$

$$\theta_{t+1,p} \sim f_2\left(\theta_2 | \theta_{t+1,1}, \theta_{t+1,2}, \dots, \theta_{t+1,p-1}\right)$$

Applications of Gibbs Sampling

- Normal model
- Regression model

Normal Model

- A set of random variable from normal distribution, knowing these variables $x_1, x_2, \dots, x_N$, use conjugate prior to calculate its mean μ, and variance σ^2

Likelihood function of the data, from the normal PDF

$$L(x|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(\sum_{i=1}^{N} -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

Prior distribution for $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$

Prior distribution for σ

$$p(\sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\nu_0 s_0^2}{2\sigma^2}\right) \sim \frac{\nu_0 s_0^2}{\chi^2_{\nu_0}}$$

Joint prior distribution $p(\mu,\sigma^2) = p(\mu) \times p(\sigma^2)$

Posterior for $\mu|\sigma^2 \sim N(\mu_1, \sigma_\mu^2)$

$$\mu_1 = \frac{\frac{\sum_1^N x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}, \qquad \sigma_\mu^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

Posterior for $\sigma^2|\mu \sim \frac{\nu_1 s_1^2}{\chi_{\nu_1}^2}$, also called $scaled - inv - \chi^2(\nu_1, s_1^2)$, or $Inv - Gamma\left(\frac{\nu_1}{2}, \frac{\nu_1 s_1^2}{2}\right)$

$$\nu_1 = \nu_0 + N, \qquad s_1^2 = \frac{\nu_0 s_0^2 + N s^2}{\nu_0 + N}$$

Regression Model

- Regression model

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$$

- Likelihood function

$$L(y, X|\beta, \sigma^2) = l(\beta, \sigma^2)$$
$$\propto \frac{1}{2\pi\sigma^N} \exp\left(-\frac{1}{2}\sum_i \frac{(y_i - X_i\beta)^2}{\sigma^2}\right)$$

- Prior distribution

$$p(\beta, \sigma^2) = p(\beta)p(\sigma^2)$$

Bayes Regression

- Prior probabilities

$$p(\beta) \propto \exp\left[\frac{-1}{2}(\beta-\bar{\beta})' A(\beta-\bar{\beta})\right]$$

$$p(\sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left[\frac{-\nu_0 s_0^2}{2\sigma^2}\right]$$

19

- Posterior of all the model parameters β, σ^2

$$\begin{aligned} &l(\beta,\sigma^2) \times prior(\beta,\sigma^2) \\ &= \frac{1}{2\pi\sigma} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - X_i\beta)^2\right) \\ &\times \exp\left(-\frac{1}{2}(\beta-\bar{\beta})' A(\beta-\bar{\beta})\right) \\ &\times (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(\frac{-\nu_0 s_0^2}{2\sigma^2}\right) \end{aligned}$$

Gibbs Sampler

Scheme: [y|X, β, σ²] [β] [σ²]

1) Draw [β | y, X, σ²]

2) Draw [σ² | y, X, β] (conditional on β!)

3) Repeat

Gibbs Sampler

- Full Conditional distribution of β, and σ^2

$$[\beta \mid y, X, \sigma^2] = N(\tilde{\beta}, (\sigma^{-2}X'X + A)^{-1})$$

with $\tilde{\beta} = (\sigma^{-2}X'X + A)^{-1}(\sigma^{-2}X'X\hat{\beta} + A\bar{\beta})$

$\hat{\beta} = (X'X)^{-1}X'y$

$$[\sigma^2 \mid y, X, \beta] = \frac{\nu_1 s_1^2}{\chi^2_{\nu_1}} \text{ with } \nu_1 = \nu_0 + n$$

$$s_1^2 = \frac{\nu_0 s_0^2 + (y - X\beta)'(y - X\beta)}{\nu_0 + n}$$

Depends on β

Summary

- MCMC basic concepts
- First MCMC: Gibbs sampler
  - Gibbs sampler and Normal model
  - Gibbs sampler for Normal regression
- Second MCMC: Metropolis-Hastings
  - Next class