More on MCMC

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

04/19/2021

Overview of posterior simulation methods

- Direct sampling
- Non-iterative indirect sampling:
  - Importance sampling
  - Rejection sampling
- Markov chain Monte Carlo sampling:
  - Metropolis-Hastings algorithm
  - Gibbs sampling
- And many more!

Connections between methods

- Accepted samples under **rejection** sampling are **direct** samples from posterior.
- **Importance** sampling is analogous to **rejection** sampling, with rejection probabilities used as weights
- **Metropolis-Hastings** sampling includes an accept/reject step similar to **rejection** sampling
- **Gibbs** sampling is a special case of **Metropolis-Hastings** sampling, in which proposal density is conditional posterior.

Combining MCMC methods

- Gibbs sampling requires direct sampling from full conditionals
- Otherwise, can combine with other sampling methods
- For example: use Gibbs sampling, with a MH step to sample from intractible conditionals
- A single MH "sub-step" is sufficient for convergence:
  - Draw $\theta_i^{(t)}$ using MH with proposal density $q(\theta_i \mid \theta_i^{(t-1)})$ and $h \propto p(\theta_i \mid \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_k^{(t-1)}, \mathbf{y})$.

Example: IQ (cont.)

- Model
  - Scores $y_{ij} \sim \text{Normal}(\theta_i, \sigma^2)$ for individuals $i = 1, \ldots, m$, trials $j = 1, \ldots, n_i$
  - IQs $\theta_i \sim \text{Normal}(\mu, 225)$ for $i = 1, \ldots, m$
- Use flat prior for μ, Gamma$(25, 1)$ for σ^2

$$p(\mu, \sigma^2) \propto (\sigma^2)^{24} e^{-\sigma^2}$$

- The full conditional for σ^2 is proportional to

$$(\sigma^2)^{24-n/2} \exp\left\{ -\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2 \right\}$$

- Not a well-known density.

Example: IQ (cont.)

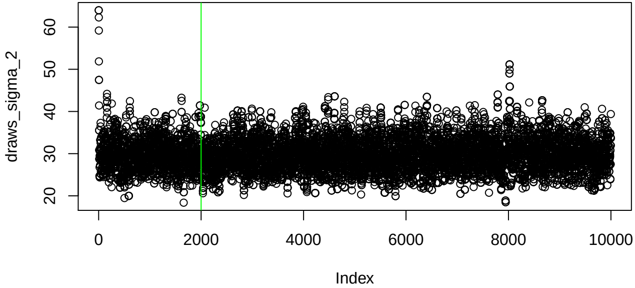
- Use Gibbs sampling, with MH sub-step for σ^2
- For $t = 1, \ldots, T$:
  - For $i = 1, \ldots, 20$ draw $\theta_i^{(t)}$ from

$$p(\theta_i|\mathbf{y}, \sigma^2, \mu) = \text{Normal}\left(\frac{\sigma^{2(t-1)}\mu^{(t-1)} + n_i\tau^2\bar{y}_i}{n_i\tau^2 + \sigma^{2(t-1)}}, \frac{\sigma^{2(t-1)}\tau^2}{n_i\tau^2 + \sigma^{2(t-1)}}\right)$$

  - Draw $\sigma^{2(t)}$ using Metropolis step
    - Draw σ^{2*} from $q(\cdot \mid \sigma^{2(t-1)}) = \text{Normal}(\sigma^{2(t-1)}, 25)$
    - Compute $r = \frac{p(\sigma^{2*}, \theta^{(t)}, \mu^{(t-1)}, \mathbf{y})}{p(\sigma^{2(t-1)}, \theta^{(t)}, \mu^{(t-1)}, \mathbf{y}))}$
    - If $r \geq 1$, set $\sigma^{2(t)} = \sigma^{2*}$;
      if $r < 1$, set $\sigma^{2(t)} = \begin{cases} \sigma^{2*} & \text{with probability } r \\ \sigma^{2(t-1)} & \text{with probability } 1 - r \end{cases}$.
  - Draw $\mu^{(t)}$ from $p(\mu \mid \mathbf{y}, \theta, \sigma^2) = \text{Normal}(\bar{\theta}^{(t-1)}, 225/m)$

Example: IQ (cont.)

- MH draws $\sigma^{2(1)}, \sigma^{2(2)}, \ldots$:

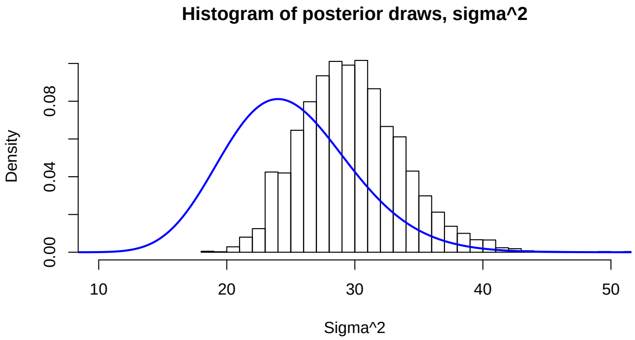


- Acceptance rate: 59%.
- Autocorrelation of draws $r = 0.765$.

http://www.ericfrazerlock.com/More_on_MCMC_Rcode1.r

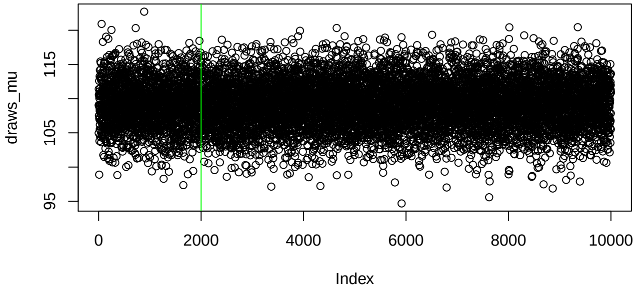
Example: IQ (cont.)

- Estimated marginal posterior density for σ^2, with prior:



Example: IQ (cont.)

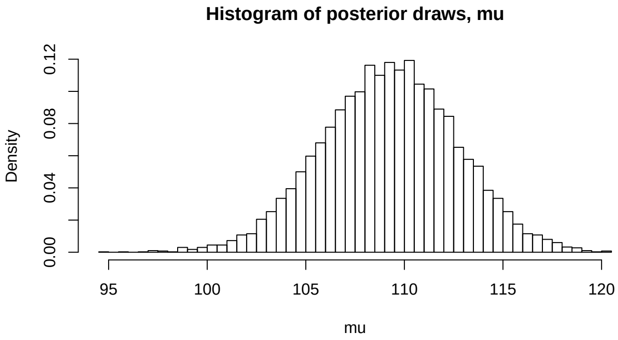
- Gibbs draws $\mu^{(1)}, \mu^{(2)}, \ldots$:



- Autocorrelation of draws $r = 0.02$.

Example: IQ (cont.)

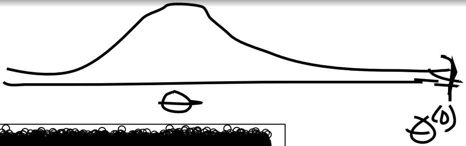
- Estimated marginal posterior density for μ:



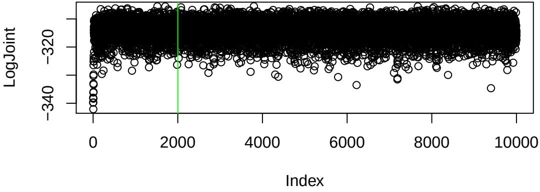
Assessing convergence

- MCMC iterations will eventually converge to their stationary distribution (the posterior)
- Can be assessed by visual inspection of *trace* plots
  - Plot of draws over the iterations for a parameter
- There should be no indication of a systematic trend, after burn-in
- The log joint density can be used as a summary
  - Consider $\log p(\theta_1^{(t)}, \ldots, \theta_k^{(t)}, \mathbf{y})$ for each iteration t
  - Would like to see this increase during convergence, then appear stationary after burn-in

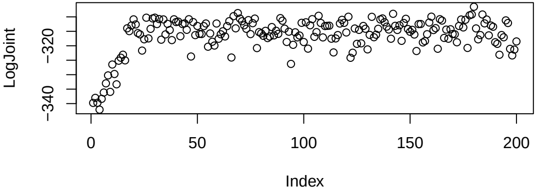
Example: IQ (cont.)



- Log-density trace plot:



- First 200 iterations



Assessing convergence: multiple initializations

- Repeat the chain in parallel from multiple initial conditions
  - Trace plots of draws should be indistinguishable after burn-in.
- Would like initializations that are well-spread over parameter space to assess robustness
  - Initializations over-dispersed with respect to posterior

$$\text{Var}(\text{initial } \theta s) > \text{Var}_{\mathbf{y}}(\theta)$$

- But don't want initial values *too* far away from posterior concentration, as this can slow convergence
- Generating initial values from prior p_θ is one approach

Assessing convergence: multiple initializations

- Run MCMC chain from m different initializations
- Let $\theta^{(t,j)}$ be the t'th iteration from j'th chain
- Consider the overall (O) and within-chain (W) variance:

$$O = \frac{1}{Nm-1}\sum_{i=1}^{N}\sum_{j=1}^{m}(\theta^{(i,j)} - \bar{\theta}^{(\cdot,\cdot)})^2$$

$$W = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{N-1}\sum_{i=1}^{N}(\theta^{(i,j)} - \bar{\theta}^{(\cdot,j)})^2\right]$$

- If chains are indistinguishable, O and W should be nearly identical.

Assessing convergence: multiple initializations

- A common diagnostic is the *scale reduction factor*

$$\sqrt{R} = \sqrt{\frac{O}{W}}.$$

- There are different, related version of $\sqrt{R}$
  - e.g., that given in (3.32) of Carlin&Louis
- First introduced by Gelman & Rubin, 1992
- Ideally R is close to 1.
- $R > 1$ implies draws vary more across chains than within chains
  - Suggests draws are still dependent on initial conditions
- Requiring $\sqrt{R} < 1.1$ for draws after burn-in is a common threshold.

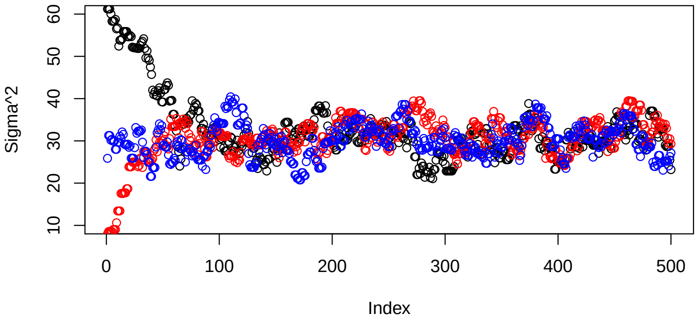
Example: IQ (cont.)

- Run previous Gibbs-Metropolis sampler for $m = 10$ different initializations
  - $T = 10000$ total draws
  - First 2000 used as burn-in: $N = 8000$
  - Use proposal density with variance 4 for σ^2 draws
- Draw $\sigma^{2(0)}$ from $Gamma(25/2, 1/2)$
  - Same expected value as prior, but more variance
- Draw $\mu^{(0)}$ from $Normal(100, 225)$.

http://www.ericfrazerlock.com/More_on_MCMC_Rcode1.r

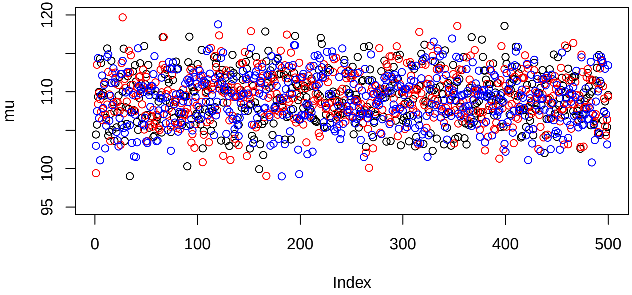
Example: IQ (cont.)

- First 500 Gibbs draws $\sigma^{2(i,1)}, \sigma^{2(i,2)}, \ldots$ for three different initializations i:



Example: IQ (cont.)

- First 500 Gibbs draws $\mu^{(i,1)}, \mu^{(i,2)}, \ldots$ for three different initializations i:



- Draws "mix" quickly

Example: IQ (cont.)

- For the first 100 draws:
  - $\sqrt{R_{\sigma^2}} = 1.324$
  - $\sqrt{R_{\mu}} = 1.000$
- For draws after burn-in, $t = 2001, .., 10000,$:
  - $\sqrt{R_{\sigma^2}} = 1.001$
  - $\sqrt{R_{\mu}} = 1.000$
- A good sign that our burn-in is sufficient!

Assessing convergence

- For multiple chains, the draws after burn-in may be combined across chains for posterior inference.

$$m \times N \text{ total draws}$$

- However, it is often preferred to simply run one long chain
  - The burn-in stage for each chain may be considered "wasteful"
- Furthermore, it is hard to be 100% confident that different initializations are well-spread over posterior support
  - Different chains may appear to converge, but to the same local mode
- There are many other convergence criteria
  - Some do not require multiple chains, and some give a single summary for all parameters
  - For an overview see Cowles & Carlin, 1996

Assessing variability due to simulation

- For $\lambda = g(\theta)$, consider the estimate for λ based on MCMC draws

$$\hat{E}(\lambda \mid \mathbf{y}) = \hat{\lambda}_N = \frac{1}{N}\sum_{t=1}^{N}\lambda^{(t)}$$

- Consider $\text{Var}(\hat{\lambda}_N)$, assuming draws are from the posterior (i.e., the MCMC has converged) but dependent.
- Define ρ_k, the autocorrelation between $\lambda^{(t)}$ and $\lambda^{(t+k)}$.
- The *effective sample size*, *ESS*, is defined by

$$ESS = N/\kappa(\lambda),$$

where

$$\kappa(\lambda) = 1 + 2\sum_{k=1}^{\infty}\rho_k(\lambda)$$

$$\mathrm{Var}\left(\frac{1}{N}\sum_{t=1}^{N}\lambda^{(t)}\right) = \frac{1}{N^2}\left[\sum_{t=1}^{N}\mathrm{Var}(\lambda^{(t)}) + 2\sum_{i<j}\mathrm{Cov}(\lambda^{(i)},\lambda^{(j)})\right]$$

$$\approx \frac{1}{N^2}\left[N S_\lambda^2 + 2\sum_{i<j}\rho_{(j-i)} S_\lambda^2\right]$$

$$= \frac{1}{N^2}\left[N S_\lambda^2 + 2\left((N-1)\rho_1 S_\lambda^2 + (N-2)\rho_2 S_\lambda^2 + \ldots\right)\right]$$

$$\approx \frac{1}{N^2}\left[N\left(S_\lambda^2 + 2(\rho_1 + \rho_2 + \ldots)S_\lambda^2\right)\right]$$

$$= \frac{1}{N} S_\lambda^2 K(\lambda)$$

$$= S_\lambda^2 / ESS$$

Assessing variability due to simulation

- The simulation variance of $\hat{\lambda}_N$ may be approximated by

$$\hat{Var}(\hat{\lambda}_N) = \frac{s_\lambda^2}{ESS}$$

where

$$s_\lambda^2 = \frac{1}{N-1}\sum_{t=1}^{N}(\lambda^{(t)} - \hat{\lambda}_N)^2.$$

- ESS can be computed by summing autocorrelations until they become negligible (say, below 0.01).
- Often autocorrelation decays exponentially: $\rho_k \approx \rho_1^k$
  - This gives

$$ESS \approx N\left(\frac{1-\rho_1}{1+\rho_1}\right)$$

$$ESS = \frac{N}{K(\lambda)} = \frac{N}{1+2\sum_{k=1}^{\infty}\rho_k(\lambda)}$$

$$\approx \frac{N}{1+2\sum_{k=1}^{\infty}\rho_1^k(\lambda)}$$

$$= \frac{N}{1+2\left(\frac{\rho_1}{1-\rho_1}\right)} = N\left(\frac{1+\rho_1}{1-\rho_1}\right)$$

Example: IQ (cont.)

- The first 10 autocorrelations for σ^2 draws are

$$\rho_1 = 0.762 \;\rho_2 = 0.592 \;\rho_3 = 0.456 \;\rho_4 = 0.354 \;\rho_5 = 0.277$$

$$\rho_6 = 0.213 \;\rho_7 = 0.163 \;\rho_8 = 0.121 \;\rho_9 = 0.091 \;\rho_{10} = 0.063$$

- The first 10 powers of ρ_1 are

$$\rho_1 = 0.762 \;\rho_1^2 = 0.581 \;\rho_1^3 = 0.443 \;\rho_1^4 = 0.338 \;\rho_1^5 = 0.258$$

$$\rho_1^6 = 0.197 \;\rho_1^7 = 0.150 \;\rho_1^8 = 0.114 \;\rho_1^9 = 0.087 \;\rho_1^{10} = 0.066$$

- Approximate

$$ESS = N\left(\frac{1-\rho_1}{1+\rho_1}\right) = 1076$$

- Estimate $\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N}\sigma^{2(t)} = 29.84$
- $\text{Var}(\hat{\sigma}^2) = s_{\sigma^2}/ESS = 0.015$