Lecture 8: Metropolis-Hastings, Gibbs and Blocking

STA702

Merlise Clyde
Duke University

Metropolis-Hastings (MH)

- Metropolis requires that the proposal distribution be symmetric
- Hastings (1970) generalizes Metropolis algorithms to allow asymmetric proposals - aka Metropolis-Hastings or MH $q(\theta^* \mid \theta^{(s)})$ does not need to be the same as $q(\theta^{(s)} \mid \theta^*)$
- propose $\theta^* \mid \theta^{(s)} \sim q(\theta^* \mid \theta^{(s)})$
- Acceptance probability

$$\min\left\{1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/q(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/q(\theta^{(s)} \mid \theta^*)}\right\}$$

- adjustment for asymmetry in acceptance ratio is key to ensuring convergence to stationary distribution!

Special cases

- Metropolis
- Independence chain
- Gibbs samplers
- Metropolis-within-Gibbs
- combinations of the above!

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$
- acceptance probability

$$\min\left\{1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/\tilde{\pi}(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/\tilde{\pi}(\theta^{(s)} \mid \theta^*)}\right\}$$

- what happens if the approximation is really accurate?
- probability of acceptance is ≈ 1
- Important caveat for convergence: tails of the posterior should be at least as heavy as the tails of the posterior (Tweedie 1994)
- Replace Gaussian by a Student-t with low degrees of freedom
- transformations of θ

Blocked Metropolis-Hastings

So far all algorithms update all of the parameters simultaneously

- convenient to break problems in to K blocks and update them separately
- $\theta = (\theta_{[1]}, \ldots, \theta_{[K]}) = (\theta_1, \ldots, \theta_p)$
- At iteration s, for $k = 1, \ldots, K$ Cycle thru blocks: (fixed order or random order)
  - propose $\theta^*_{[k]} \sim q_k(\theta_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})$
  - set $\theta^{(s)}_{[k]} = \theta^*_{[k]}$ with probability

$$\min\left\{1, \frac{\pi(\theta^{(s)}_{[<k]}, \theta^*_{[k]}, \theta^{(s-1)}_{[>k]})\mathcal{L}(\theta^{(s)}_{[<k]}, \theta^*_{[k]}, \theta^{(s-1)}_{[>k]})/q_k(\theta^*_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})}{\pi(\theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[k]}, \theta^{(s-1)}_{[>k]})\mathcal{L}(\theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[k]}, \theta^{(s-1)}_{[>k]})/q_k(\theta^{(s-1)}_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})}\right.$$

Gibbs Sampler

- The Gibbs Sampler is special case of Blocked MH
- proposal distribution q_k for the kth block is the **full conditional** distribution for $\theta_{[k]}$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) = \frac{\pi(\theta_{[k]}, \theta_{[-k]} \mid y)}{\pi(\theta_{[-k]} \mid y))} \propto \pi(\theta_{[k]}, \theta_{[-k]} \mid y)$$
$$\propto \mathcal{L}(\theta_{[k]}, \theta_{[-k]})\pi(\theta_{[k]}, \theta_{[-k]})$$

- Acceptance probability

$$\min\left\{1, \frac{\pi(\theta^{(s)}_{[<k]}, \theta^*_{[k]}, \theta^{(s-1)}_{[>k]})\mathcal{L}(\theta^{(s)}_{[<k]}, \theta^*_{[k]}, \theta^{(s-1)}_{[>k]})/q_k(\theta^*_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})}{\pi(\theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[k]}, \theta^{(s-1)}_{[>k]})\mathcal{L}(\theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[k]}, \theta^{(s-1)}_{[>k]})/q_k(\theta^{(s-1)}_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})}\right.$$

- Simplifies so that acceptance probability is always 1!
- even though joint distribution is messy, full conditionals may be (conditionally) conjugate and easy to sample from!

Univariate Normal Example

Model

$$\begin{aligned} Y_i \mid \mu, \sigma^2 &\overset{iid}{\sim} \mathsf{N}(\mu, 1/\phi) \\ \mu &\sim \mathsf{N}(\mu_0, 1/\tau_0) \\ \phi &\sim \mathsf{Gamma}(a/2, b/2) \end{aligned}$$

- Joint prior is a product of independent Normal-Gamma
- Is $\pi(\mu, \phi \mid y_1, \ldots, y_n)$ also a Normal-Gamma family?

Full Conditional for the Mean

The full conditional distributions $\mu \mid \phi, y_1, \ldots, y_n$

$$\mu \mid \phi, y_1, \ldots, y_n \sim \mathsf{N}(\hat{\mu}, 1/\tau_n)$$
$$\hat{\mu} = \frac{\tau_0 \mu_0 + n\phi\bar{y}}{\tau_0 + n\phi}$$
$$\tau_n = \tau_0 + n\phi$$

Full Conditional for the Precision

- Full conditional for ϕ

$$\begin{aligned} \phi \mid \mu, y_1, \ldots, y_n &\sim \textsf{Gamma}(a_n/2, b_n/2) \\ a_n &= a + n \\ b_n &= b + \sum_i (y_i - \mu)^2 \end{aligned}$$

$$\textsf{E}[\phi \mid \mu, y_1, \ldots, y_n] = \frac{(a + n)/2}{(b + \sum_i (y_i - \mu)^2)/2}$$

- What happens with a non-informative prior i.e $a = b = \epsilon$ as $\epsilon \to 0$?

⚠ Proper full conditionals with improper priors do not ensure proper joint posterior!

Normal Linear Regression Example

- Model

$$
\begin{aligned}
Y_i \mid \beta, \phi &\stackrel{iid}{\sim} \textsf{N}(x_i^T\beta, 1/\phi) \\
Y \mid \beta, \phi &\sim \textsf{N}(X\beta, \phi^{-1} I_n) \\
\beta &\sim \textsf{N}(b_0, \Phi_0^{-1}) \\
\phi &\sim \textsf{N}(v_0/2, s_0/2)
\end{aligned}
$$

- x_i is a $p \times 1$ vector of predictors and X is $n \times p$ matrix
- β is a $p \times 1$ vector of coefficients
- Φ_0 is a $p \times p$ prior precision matrix
- Multivariate Normal density for β

$$
\pi(\beta \mid b_0, \Phi_0) = \frac{|\Phi_0|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\beta - b_0)^T \Phi_0 (\beta - b_0)\right\}
$$

Full Conditional for β

$$
\begin{aligned}
\beta \mid \phi, y_1, \ldots, y_n &\sim \mathsf{N}(b_n, \Phi_n^{-1}) \\
b_n &= (\Phi_0 + \phi X^T X)^{-1}(\Phi_0 b_0 + \phi X^T X \hat{\beta}) \\
\Phi_n &= \Phi_0 + \phi X^T X
\end{aligned}
$$

Derivation continued

Full Conditional for ϕ

$$\phi \mid \beta, y_1, \ldots, y_n \sim \textsf{Gamma}((v_0 + n)/2, (s_0 + \sum_i (y_i - x_i^T \beta)))$$

Choice of Prior Precision

- Non-Informative $\Phi_0 \to 0$
- Formal Posterior given ϕ

$$\beta \mid \phi, y_1, \ldots, y_n \sim \mathsf{N}(\hat{\beta}, \phi^{-1}(X^T X)^{-1})$$

- needs $X^T X$ to be full rank for distribution to be unique

Invariance and Choice of Mean/Precision

- the model in vector form

$$Y \sim \mathsf{N}_n(X\beta, \phi^{-1} I_n)$$

- What if we transform the X matrix by $\tilde{X} = XH$ where H is $p \times p$ and invertible
- obtain the posterior for $\tilde{\beta}$ using Y and $\tilde{X}$

$$Y \sim \mathsf{N}_n(\tilde{X}\tilde{\beta}, \phi^{-1} I_n)$$

- since $\tilde{X}\tilde{\beta} = XH\tilde{\beta} = X\beta$ invariance suggests that the posterior for β and $H\tilde{\beta}$ should be the same
- or the posterior of $H^{-1}\beta$ and $\tilde{\beta}$ should be the same
- with some linear algebra we can show that this is true if $b_0 = 0$ and Φ_0 is $kX^T X$ for some k (show!)

Zellner's g-prior

- Popular choice is to take $k = \phi/g$ which is a special case of Zellner's g-prior

$$\beta \mid \phi, g \sim \mathsf{N}\left(0, \frac{g}{\phi}(X^T X)^{-1}\right)$$

- Full conditional

$$\beta \mid \phi, g \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\beta}, \frac{1}{\phi}\frac{g}{1+g}(X^T X)^{-1}\right)$$

- one parameter g controls shrinkage
- if $\phi \sim \mathsf{Gamma}(v_0/2, s_0/2)$ then posterior is

$$\phi \mid y_1, \ldots, y_n \sim \mathsf{Gamma}(v_n/2, s_n/2)$$

- Conjugate so we could skip Gibbs sampling and sample directly from gamma and then conditional normal!

Ridge Regression

- If X^TX is nearly singular, certain elements of β or (linear combinations of β) may have huge variances under the g-prior (or flat prior) as the MLEs are highly unstable!
- **Ridge regression** protects against the explosion of variances and ill-conditioning with the conjugate priors:

$$\beta \mid \phi \sim \mathsf{N}(0, \frac{1}{\phi\lambda} I_p)$$

- Posterior for β (conjugate case)

$$\beta \mid \phi, \lambda, y_1, \ldots, y_n \sim \mathsf{N}\left((\lambda I_p + X^TX)^{-1} X^TY, \frac{1}{\phi}(\lambda I_p + X^TX)^{-1}\right)$$

Bayes Regression

- Posterior mean (or mode) given λ is biased, but can show that there **always** is a value of λ where the frequentist's expected squared error loss is smaller for the Ridge estimator than MLE!
- related to penalized maximum likelihood estimation
- Choice of λ
- Bayes Regression and choice of Φ_0 in general is a very important problem and provides the foundation for many variations on shrinkage estimators, variable selection, hierarchical models, nonparametric regression and more!
- Be sure that you can derive the full conditional posteriors for β and ϕ as well as the joint posterior in the conjugate case!

Comments

- Why don't we treat each individual β_j as a separate block?
- Gibbs always accepts, but can mix slowly if parameters in different blocks are highly correlated!
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)
- Collapse the sampler by integrating out as many parameters as possible (as long as resulting sampler has good mixing)
- can use Gibbs steps and (adaptive) Metropolis Hastings steps together
- Introduce latent variables (data augmentation) to allow Gibbs steps (Next class)