Markov Chain Monte Carlo (2/2)

MCMC

- Gibbs Sampler
  - Help to decompose big complex problem (find joint distribution) into small simple problems (use conditional distributions)
  - Limitation: still needs to know the conditional distributions, therefore, only conjugate
- Metropolis-Hastings
  - Can solve any model!

Reversibility vs. Stationarity

- Reversibility with respect to ω

$$\omega_i p_{ij} = \omega_j p_{ji}$$

- Summing both sides over i

$$\sum_i \omega_i p_{ij} = \sum_i \omega_j p_{ji} = \omega_j \sum_i p_{ji} = \omega_j$$

that is $\omega P = \omega$, which is stationary

This provides a method to find the stationary distribution of the chain

Metropolis Methods

Goal:

construct a Markov Chain whose invariant (stationary) distribution is the posterior, using only non-normalized posterior.

Metropolis idea:

given a chain which is easy to sample from, modify to have π as its invariant distribution

[similar to accept/reject sampling – sample from a proposal, accept/reject to obtain desired distribution]

Discrete Case: Metropolis-Hastings algorithm

i. start with a chain defined by transition matrix Q.

ii. modify to new chain with π as invariant distribution.

iii. require only un-normalized posterior

How? Use the principle of time reversibility wrt π

Discrete Case: Metropolis-Hastings algorithm

$q_0 = q^i \quad (\text{start in state } i)$

draw state j with prob $\left(q_{i,1}, \ldots, q_{id}\right)$

compute $a = \min\left\{1, \dfrac{p_j\, q_{j,i}}{p_i\, q_{i,j}}\right\}$

with prob a $\quad q_1 = q^j \ (\text{move})$

else $\quad q_1 = q^i \ (\text{stay})$

Note: with prob 1-α,
this chain will
repeat!!

Discrete Case: Metropolis-Hastings algorithm

why repeat?

$$\text{if } \pi_i q_{i,j} > \pi_j q_{j,i} \Rightarrow \alpha < 1$$

“too many” transitions from i to j

“not enough” transitions from j to i

if at state i, repeat i to lower number of transitions.

if at state j, always move to i!

Time reversible wrt π

$$p_{ij} = q_{ij}\,\alpha(i,j)$$

generating candidate j given i

acceptance probability

$$\pi_i p_{ij} = \pi_i q_{ij} \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\} = \min\{\pi_i q_{ij}, \pi_j q_{ji}\}$$

$$\pi_j p_{ji} = \min\{\pi_j q_{ji}, \pi_i q_{ij}\}$$

$$\Rightarrow \pi_i p_{i,j} = \pi_j p_{j,i}$$

Metropolis-Hastings algorithm example

$$\pi = \begin{bmatrix} 1/3 & 2/3 \end{bmatrix} \qquad q_{ij} = 1/2 \qquad Q = \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix}$$

$$p_{12} = .5\min\left\{1, \frac{2/3}{1/3}\right\} = .5(1) = .5$$

$$p_{21} = .5\min\left\{1, \frac{1/3}{2/3}\right\} = .5(.5) = .25 \quad P = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

check: does π1p12 = π2p21? does πP=π? (yes!)

9

Continuous Metropolis-Hastings

discrete: i ➔ j

continuous: $\theta \to \vartheta$

Q is a the proposal Markov chain. $q(\theta,\vartheta)$ is the kernel.
π is the desired stationary distribution.

1. Generate $\vartheta \sim q(\theta,\vartheta)$

2. $\alpha(\theta,\vartheta) = \min\left\{1, \frac{\pi(\vartheta)\,q(\vartheta,\theta)}{\pi(\theta)\,q(\theta,\vartheta)}\right\}$

3. With prob α, move to ϑ, else stay at θ

How to Find Proposal q?

- Ideally, we want the proposal distribution $q(\theta, \vartheta)$to have a fatter tail than the target distribution $\pi(\theta)$
  - Method 1: Random-walk proposal function
  - Method 2: Independence chain

Random Walk MH

- Random-walk proposal function

$$\vartheta = \theta + \epsilon$$

$$q(\theta, \vartheta) = q_\epsilon(\vartheta - \theta) \sim N(0, s^2 I)$$

- Random-walk Metropolis Chain

```
Start with $\theta_0$
Draw $\vartheta = \theta + \epsilon$, $\epsilon \sim N(0, s^2 I)$
Compute $\alpha = \min\{1, \pi(\vartheta)/\pi(\theta)\}$
With probability $\alpha$, take the proposal $\theta_1 = \vartheta$
With probability $1 - \alpha$, stay $\theta_1 = \theta_0$
Repeat
```

Independence chain

Let $q(\theta,\vartheta) = q_{imp}(\vartheta)$

“ind of current location.
“imp” for importance
function

$$\text{Then } \alpha(\theta,\vartheta) = \min\left\{1, \frac{\pi(\vartheta)\, q_{imp}(\theta)}{\pi(\theta)\, q_{imp}(\vartheta)}\right\}$$

$$= \min\left\{1, \frac{\pi(\vartheta)/q_{imp}(\vartheta)}{\pi(\theta)/q_{imp}(\theta)}\right\}$$

qimp() should have fatter tails than π to avoid the need to reject draws to build up tail mass.

Independence chain

if q is an excellent approximation to π,

$$\frac{\pi(\theta)}{q_{imp}(\theta)} \approx \text{constant}$$

α will be approximately 1!

how does it work?

if π has *more* mass (relative to q) at ϕ than at θ, move to ϕ with prob 1.

if π has *less* mass (relative to q) at ϕ than at θ, stay with some prob > 0 to build up mass

Independence chain

"important" that qimp have fatter tails.

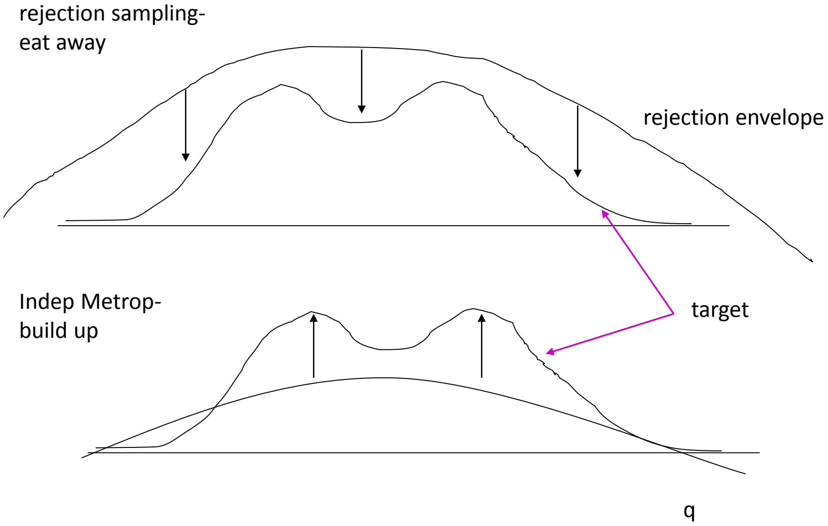
If

$$\pi(\theta) \le M q_{imp}$$

then independence Metropolis is *uniformly* ergodic

Ind. Metropolis works well in low dimensions

Independence Metro vs. Rejection Sampling



Independence vs. RW Chains

Independence Chains:

requires a good approximation to posterior
(similar to Importance Sampling)
implies some sort of optimizer
more efficient than RW

RW Chains:

will explore parameter space – no location required!
for low dimensions will work even with “dumb” choices of increment Cov matrix
may not work well in high dimensional spaces unless increment Cov closely approximates posterior

Choosing a step size for the RW chain

At θ, draw $\varepsilon \sim q$ independent of θ.
candidate $= \theta + \varepsilon$

ε very small leads to small steps, higher acceptance, higher autocorrelation.

ε very large leads to large steps, lower acceptance, higher autocorrelation.

Pick $\varepsilon \sim N(0,s^2\Sigma)$, choosing s to maximize information content.

Choosing a step size for the RW chain

Choice of Σ:

I

Asymptotic Var-Cov for Posterior or Likelihood

Run chain with I, then use cov matrix of draws

Choice of scaling constant (s):

Method 1: get the “right” acceptance rate (30-50%)

Method 2:

$$s = 2.93 \Big/ \sqrt{d = \dim(\text{state space})}$$

Applications to MNL Model

- MNL model, likelihood function

$$l(\beta) = \prod_i \prod_j \frac{\exp\left(X_{ij}\beta_j\right)^{y_{ij}}}{\sum_k \exp(X_{ik}\beta_k)^{y_{ik}}}$$

i- for data points
j,k – for alternatives

- Prior for the model parameters

$$\beta \sim MVN(\beta_0, \Sigma_0)$$

$$\exp\left(-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)\right)$$

- Posterior: multiply the two equations above, we get $\pi(\beta)$
  - Unknown distribution, no conjugacy
  - We try to establish a Markov Chain so that the stationary distribution is $\pi(\beta)$
  - We cannot find the transition matrix easily, we use Metropolis-Hastings
    - Try a proposal transition matrix
    - Adjust the chain based on MH algorithm

Logit model-Hessian

Both Indep and RW Metropolis chains rely on an asymptotic approximation to the posterior

$$p\left(b\middle|X,y\right)\,\mu\,\left|H\right|^{1/2}\exp\left\{\tfrac{1}{2}\left(b-\hat{b}\right)'H\left(b-\hat{b}\right)\right\}$$

Method 1: we can use the expected sample information matrix:

$$H=-E\left[\frac{\partial^2\log\ell}{\partial b\,\partial b'}\right]=\sum_i X_iA_iX_i'$$

$$X=\begin{bmatrix}X_1\\ \vdots\\ X_n\end{bmatrix};\quad A_i=\mathrm{Diag}\left(p_i\right)-p_ip_i'$$

Method 2: use the MLE estimate and minus Hessian

Logit model MCMC Algorithms

1. Pick an arbitrary starting value β^{old}

2. Generate candidate realization:
   random walk chain: $\beta^{cand}=\beta^{old}+\varepsilon;\quad \varepsilon\sim N\left(0,s^2H^{-1}\right)$

3. Accept β^{cand} with probability α

$$\alpha=\min\left\{1,\frac{l\left(\beta^{cand}|y,X\right)p\left(\beta^{cand}\right)}{l\left(\beta^{old}|y,X\right)p\left(\beta^{old}\right)}\times\frac{q\left(\beta^{cand},\beta^{old}\right)}{q\left(\beta^{old},\beta^{cand}\right)}\right\}$$

4. Repeat