The Metropolis-Hastings Algorithm

Constructs a Markov chain w/ given stationary (and limiting) dist. π where π is the target, usually known up to a constant of proportionality.
That is, only need $h(x)$ where $h(x)/c = \pi(x)$, c unknown and intractable. M-chain: $X_1, X_2, \dots$

$\hat{\mu}_n = \sum_{i=1}^{n} g(X_i)/n$ is Markov chain Monte Carlo approx. to $\mu = E_\pi\{g(x)\}$

Under some conditions, SLLN; more conditions, CLT.

Let support of $\pi(x)$ be Ω.

All-at-once M-H:

Start w/ $X_0 = x \in \Omega$ initial state

For $n = 0, 1, 2, \dots$ if $X_n = x$, X_{n+1} is generated as follows:

(1) Generate a candidate ("proposal") $y^* \sim q(y|x)$ (often denoted $q(x,y)$)

(2) Set $X_{n+1} = y^*$, "accept" proposal y^*, w/ probability

$$\alpha(x, y^*) = \begin{cases} \min\left(\dfrac{h(y)}{h(x)} \dfrac{q(y,x)}{q(x,y)}, 1\right) & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{else} \end{cases}$$

else "reject" proposal and set $X_{n+1} = x$.

For q, need:

(i) $q(x,y) = 0 \Rightarrow q(y,x) = 0 \quad \forall x, y \in \Omega$

(ii) $q(x,y)$ is transition kernel of irreducible M-chain on Ω.

E.g. Model: $Y_i | \theta \sim N(\theta, 1)$ condtl. indep., $i=1,\dots,n$

$\theta \sim \text{Log-}t\,(\mu, \sigma, r)$

What is $E_\pi[\theta]$ where $\pi(\theta | \underline{Y})$ is posterior distr. of θ.

$$\pi(\theta | \underline{Y}) \propto \mathcal{L}(\underline{Y} | \theta)\, p(\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(Y_i - \theta)^2\right\} \frac{1}{\theta}\left[1 + \frac{1}{r}\left(\frac{\log\theta - \mu}{\sigma}\right)\right]^{-(r+1)/2}$$

$$\propto \frac{1}{\theta} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Y_i - \theta)^2\right\} \times \left\{1 + \frac{1}{r}\left(\frac{\log\theta - \mu}{\sigma}\right)\right\}^{-(r+1)/2}$$

MCMC algorithm:

Let $h(\theta) = \frac{1}{\theta} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Y_i - \theta)^2\right\} \left\{1 + \frac{1}{r}\left(\frac{\log\theta - \mu}{\sigma}\right)\right\}^{-(r+1)/2}$

Need proposal q.

The Metropolis algorithm ('random walk' Metropolis-Hastings)

M-H algorithm where $q(x,y) = q(y,x)$ for all x, y.

That is, M-H algorithm w/ <u>symmetric proposal</u>, so

acceptance prob. $\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{q(y,x)}{q(x,y)}\right\}$

$= \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$

E.g. $q(x,y)$ is a normal density centered at x.

Propose new value $y^* \sim q(x, \cdot)$ so $y^* \sim N(x, \tau^2)$

Variance τ^2 is a tuning parameter: affects how well the algorithm performs.

τ^2 too big: candidates generated far from current value, maybe in tails $\Rightarrow$ low prob. of being accepted.

τ^2 too small: proposals/candidates accepted often but too close to previous value $\Rightarrow$ chain explores state space very slowly and high autocorrelations across sampled values (large variance/M.C. error)

Return to our example. Want to simulate from $\pi(\theta|\underline{y})$.

Suppose we use Metropolis algorithm.
$q(\theta, \theta^*)$ is $N(\theta, \tau^2)$. Symmetric since $q(\theta,\theta^*) = q(\theta^*, \theta)$.

Algorithm: When current value of M.chain, say, $\theta^{(n)} = \theta$,
propose new value $\theta^* \sim N(\theta, \tau^2)$.
Accept θ^* w/ prob. $\alpha(\theta, \theta^*) = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta)}\right\}$
$= \min\left\{1, \frac{h(\theta^*)}{h(\theta)}\right\}$

M-H recipe:
Start M.C. at $\theta^{(0)} = c$ for some $c > 0$.

For $i = 1, \dots, n$
Propose $\theta^* \sim N(\theta^{(i-1)}, \tau^2)$
Accept θ^* w/ probability $\alpha(\theta^{(i-1)}, \theta^*) = \min\left\{1, \frac{h(\theta^*)}{h(\theta^{(i-1)})}\right\}$
i.e., $\theta^{(i)} = \theta^*$
else reject, i.e., $\theta^{(i)} = \theta^{(i-1)}$.

For any expectation that exists, $\mu = E_\pi\{g(\theta)\}$ can obtain an estimate $\hat{\mu}_n = \frac{\sum_{i=1}^{n} g(\theta^{(i)})}{n}$.

$\therefore$ $\hat{\mu}_n \xrightarrow{P} \mu$ as $n \to \infty$

$\underline{\text{All-at-once M-H (A-MH)}}$: given current state $X_n = x_n$, propose $\underbrace{\text{a single}}_{\text{"joint"}}$ update to obtain X_{n+1}.

If target is high-dimensional, may be very hard to find appropriate proposal q.

Note: q completely specifies transition kernel K (generalization of transition prob. matrix) of Markov chain.

$\underline{\text{Variable-at-a-time M-H (V-MH)}}$:

- Apply M-H update to components / sub-blocks: only need to construct low-dimensional M-H updates

- Instead of working w/ joint distr. π, work w/ "full conditional distributions" = π(component | all other components)

Variable-at-a-time M-H: (for multivariate distr.)

Main idea: suppose target distr. has K components.

'All-at-once' M-H alg. (already discussed)

$$\underset{\sim}{X}^{(n)} \to \underset{\sim}{X}^{(n+1)}$$

$$\begin{pmatrix} X_1^{(n)} \\ X_2^{(n)} \\ \vdots \\ X_K^{(n)} \end{pmatrix} \xrightarrow[\text{via M-H step}]{\text{update}} \begin{pmatrix} X_1^{(n+1)} \\ X_2^{(n+1)} \\ \vdots \\ X_K^{(n+1)} \end{pmatrix}$$

Variable-at-a-time M-H: n^{th} update done in small steps

$$\begin{pmatrix} X_1^{(n)} \\ X_2^{(n)} \\ \vdots \\ X_K^{(n)} \end{pmatrix} \xrightarrow[\text{components } (X_1^{(n)} \mid X_2^{(n)}, \dots, X_K^{(n)})]{\text{M-H update, conditioning on all other}} \begin{pmatrix} X_1^{(n+1)} \\ X_2^{(n)} \\ \vdots \\ X_K^{(n)} \end{pmatrix} \xrightarrow[\text{all other components } (X_2^{(n)} \mid X_1^{(n+1)}, X_3^{(n)}, \dots, X_K^{(n)})]{\text{M-H update, given}} \begin{pmatrix} X_1^{(n+1)} \\ X_2^{(n+1)} \\ X_3^{(n)} \\ \vdots \\ X_K^{(n)} \end{pmatrix}$$

$$\cdots \quad \begin{pmatrix} X_1^{(n+1)} \\ \vdots \\ X_{K-1}^{(n+1)} \\ X_K^{(n)} \end{pmatrix} \xrightarrow{\text{MH update}} \begin{pmatrix} X_1^{(n+1)} \\ \vdots \\ X_K^{(n+1)} \end{pmatrix}$$

Note: each component may itself be multidimensional.

Example: suppose $\underset{\sim}{x} = (x_1, x_2, x_3)$ where each block may ~~also~~ be multidimensional.

full condtl.

Need: $K_{1|(2,3)}$ w/ stationary distr. $\pi_{1|(2,3)}(x_1 | x_2, x_3)$

$K_{2|(1,3)}$ " " " $\pi_{2|(1,3)}(x_2 | x_1, x_3)$

$K_{3|(1,2)}$ " " " $\pi_{3|(1,2)}(x_3 | x_1, x_2)$

Construct $K_{1|2,3}$, $K_{2|1,3}$, $K_{3|1,2}$ by using M-H alg. w/ proposals q_1, q_2, q_3 respectively

If full condtl. distr. is standard distr., can directly sample from it. For e.g. if $\pi_{1|2,3}$ = Normal density simulate update for x_1 from a normal.

→ Skip below

If current state $= \underset{\sim}{X}^{(n)} = (x_1^{(n)}, x_2^{(n)}, x_3^{(n)})$ produce next state $= \underset{\sim}{X}^{(n+1)} = (x_1^{(n+1)}, x_2^{(n+1)}, x_3^{(n+1)})$ in 3 steps:

① Propose $x_1^* \sim q_1(x_1^{(n)}, x_1^* | x_2^{(n)}, x_3^{(n)})$

Accept x_1^*, ie, set $x_1^{(n+1)} = x_1^*$ w/ prob.

$$\alpha(x_1^{(n)}, x_1^* | x_2^{(n)}, x_3^{(n)}) = \min\left\{1, \frac{\pi_{1|2,3}(x_1^* | x_2^{(n)}, x_3^{(n)})\; q(x_1^{(n)}, x_1^* | x_2^{(n)}, x_3^{(n)})}{\pi_{1|2,3}(x_1^{(n)} | x_2^{(n)}, x_3^{(n)})\; q(x_1^*, x_1^{(n)} | x_2^{(n)}, x_3^{(n)})}\right\}$$

else set $x_1^{(n+1)} = x_1^{(n)}$ (reject x^*).

② Propose $x_2^* \sim q_2(x_2^{(n)}, x_2^* \mid x_1^{(n+1)}, x_3^{(n)})$

(note updated value)

(Accept) Set $x_2^{(n+1)} = x_2^*$ w/ prob. $\alpha(x_2^{(n)}, y_2 \mid x_1^{(n+1)}, x_3^{(n)})$

else (Reject): $x_2^{(n+1)} = x_2^{(n)}$.

③ Propose $y_3 \sim q_3(x_3^{(n)}, x_3^* \mid x_1^{(n+1)}, x_2^{(n+1)})$

(Accept): set $x_3^{(n+1)} = x_3^*$ w/ prob. $\alpha(x_3^{(n)}, x_3^* \mid x_1^{(n+1)}, x_2^{(n+1)})$

(Reject) else $x_3^{(n+1)} = x_3^{(n)}$.

The Markov chain constructed by this algorithm is Harris-ergodic w/ stationary distribution π

Simple example : Poi-Gamma model (C & Louis pg. 143)

$Y_i | \theta_i \sim Poi(\theta_i t_i)$ condtl. indep. $i = 1, .., k$

Prior $\theta_i | \beta \sim G(\alpha, \beta)$

$t_1, \ldots, t_k$ known ; a known.

Hyperprior $\beta \sim G(c, d)$. c, d known.

Inference based on posteria distribution

$$\pi(\underset{\sim}{\theta}, \beta | \underset{\sim}{Y}) \propto \mathcal{L}(\underset{\sim}{Y} | \underset{\sim}{\theta}) \prod_{i=1}^{n} f_1(\theta_i | \beta) \; f_2(\beta)$$

$$= \left\{ \prod_{i=1}^{k} \frac{(\theta_i t_i)^{Y_i} e^{-\theta_i t_i}}{Y_i!} \right\} \times \left\{ \prod_{i=1}^{k} \frac{1}{\Gamma(a) \beta^{a}} \theta_i^{a-1} e^{-\theta_i / \beta} \right\} \times \frac{1}{\Gamma(c) d^{c}} \beta^{c-1} e^{-\beta/d}$$

constants

$$\propto \left\{ \prod_{i=1}^{k} (\theta_i t_i)^{Y_i} e^{-\sum_i^{k} \theta_i t_i} \right\} \frac{\left(\prod_{i=1}^{k} \theta_i^{a-1} \right) e^{-\sum \theta_i / \beta}}{} \beta^{c-1-ka} e^{-\beta/d}$$

Full condtl. : ~~only keep~~

$$\pi(\theta_i | \beta, \underset{\sim}{Y}, \underset{\sim}{\theta}_{-i}) \propto (\theta_i t_i)^{Y_i} e^{-\theta_i t_i} \theta_i^{a-1} e^{-\theta_i / \beta}$$

$$\propto \theta_i^{Y_i + a - 1} e^{-\theta_i (t_i + \frac{1}{\beta})} \quad \cancel{\propto \text{Gamma}(Y_i + a}$$

$$= \theta_i^{(Y_i + a) - 1} e^{-\theta_i / (t_i + \frac{1}{\beta})^{-1}} \propto \text{Gamma}\left(Y_i + a, \left(t_i + \frac{1}{\beta}\right)^{-1}\right)$$

Recognized density! Simulate from Gamma directly, a $\boxed{\text{Gibbs' update}}$.
In fact priors that result in same type of (posterior) distr. for a given likelihood are called 'conjugate' priors.

Aside: conjugate priors are often used to make the algorithm simpler to implement (Gibbs updates).

But Gibbs update $\not\Rightarrow$ more efficient than M-H update!

Hence, not crucial to use conjugate priors unless they are reasonable expression of prior info.

$$\pi(\beta \mid \underset{\sim}{\theta}, \underset{\sim}{Y}) \propto e^{-\sum_{i=1}^{k}\theta_i/\beta} \, \beta^{c-1-a} \, e^{-\beta/d} = h(\beta \mid \underset{\sim}{\theta}, \underset{\sim}{Y}), \text{ say}$$

Not recognizable density.

M-H algorithm/update: e.g. simplest one

Propose $\beta^* \sim N(\beta_{\text{current}}, \underbrace{\tau^2}_{\uparrow \text{ tuning parameter}})$

Accept-reject via M-H prob.

So an M-H algorithm for $\pi(\underset{\sim}{\theta}, \beta \mid Y)$ is:

1) Start M.C. at $(\underset{\sim}{\theta}^{(1)}, \beta^{(1)})$ initial values

any value that is reasonably likely under π is fine.

2) Nth update of each θ_i for $i=1,\dots,k$ is according to

$$\pi(\theta_i \mid \underbrace{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k, \beta, \underset{\sim}{Y}}_{\text{most recent values}})$$

For this simple example above is just $\pi(\theta_i \mid \beta, \underset{\sim}{Y})$

$= \text{Gamma}\left(Y_i + a, \left[t_i + \frac{1}{\beta}\right]^{-1}\right)$ by condtl. indep. of θ_i's given β.

Sample $\theta_i^{(n)} \sim \text{Gamma}\left(Y_i + a, \left(t_i + \frac{1}{\beta^{(n)}}\right)^{-1}\right)$ for $i=1,\dots,k$.

'Gibbs' update: always accept

3) Nth update of β is according to $\pi(\beta \mid \underset{\sim}{\theta}, \underset{\sim}{Y})$.

Propose $\beta^* \sim N(\beta^{(n)}, \tau^2)$

Accept w/ prob. $\alpha(\beta, \beta^* \mid \underset{\sim}{\theta}, \underset{\sim}{Y}) = \min\left\{1, \frac{h(\beta^* \mid \underset{\sim}{\theta}, \underset{\sim}{Y}) \, \cancel{q(\beta^*, \beta)}}{h(\beta^{(n)} \mid \underset{\sim}{\theta}, \underset{\sim}{Y}) \, \cancel{q(\beta, \beta^*)}}\right\}$

4) Return to step (2).

M.C. produced has stationary distr. $\pi(\underset{\sim}{\theta}, \beta \mid \underset{\sim}{Y})$ and is Harris ergodic.

Some other options:

⇒ depend on current value of β

① Propose $\beta^* \sim \text{Gamma}(\gamma_1(\beta), \gamma_2(\beta))$

w/ mean = current value and variance τ^2.

so, $\gamma_1(\beta)\gamma_2(\beta) = \beta$ and $\gamma_1(\beta)\gamma_2(\beta) = \tau^2$

$\Rightarrow \gamma_2(\beta) = \tau^2/\beta$ and $\gamma_1(\beta) = \beta/\gamma_2(\beta) = \beta^2/\tau^2$.

$q(\beta, \beta^*) \neq q(\beta^*, \beta)$ so M-H accept prob.

$$= \alpha(\beta, \beta^*) = \min\left\{1, \frac{h(\beta^* \mid \underset{\sim}{\theta}, \underset{\sim}{Y})}{h(\beta \mid \underset{\sim}{\theta}, \underset{\sim}{Y})} \frac{q(\beta^*, \beta)}{q(\beta, \beta^*)}\right\}$$

where $q(\beta, \beta^*) = \text{Gamma}(\gamma_1(\beta), \gamma_2(\beta))$ pdf evaluated at β^*.

② Log-transform β, i.e., set $\psi = \log\beta \in (-\infty, \infty)$

Now use random-walk M-H update to sample from $\psi \mid \underset{\sim}{\theta}, Y$.

Can transform to get β draws, i.e., $\beta = \exp(\psi)$.

③ Laplace approx. for $\pi(\beta \mid \underset{\sim}{\theta}^{(n)}, \underset{\sim}{Y})$ as proposal $q(\beta, \beta^*)$.

Some basic M.C. theory for discrete time, cntns. state spaces. (Borrowing from Jones & Hobert, 2001)

M.C.: $X_0, X_1, X_2, \ldots \qquad X_i \in \Omega$

Discrete state space: t.p.m. $\{P_{ij}\}$ where $P_{ij} = Pr$(move to state j from state i) $= P(X_n = j \mid X_{n-1} = i) \quad i, j \in \Omega$

Cntns. state space: transition density (more generally, the 'transition kernel') is a condtl pdf, $K(x,y) = k(y|x)$ s.t.

$$P(X_n \in A \mid X_{n-1} = x) = \int_A k(y|x)\,dy \quad \forall x \in \Omega, \text{ all intervals } A.$$

Technically: $\forall x \in \Omega, \ \forall A \in \mathcal{B}(\Omega)$

Borel σ-algebra generated by Ω (collection of subsets of Ω)

$K^n(x,y)$ is n-step transition kernel

$$P(X_{i+n} \in A \mid X_i = x) = \int_A k^n(y|x)\,dx$$

This is an M.C. so

$$P(X_n \in A \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots) = P(X_n \in A \mid X_{n-1} = x_{n-1})$$

$$= \int_A k(y|x_{n-1})\,dy$$

E.g. of M.C. on contns. state space: AR(1) model

$$X_n = \theta X_{n-1} + \varepsilon_n \qquad \theta \in \mathbb{R}$$

$$\varepsilon_1, \varepsilon_2, \ldots \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

$$P(X_n \in A \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots)$$
$$= P(X_n \in A \mid X_{n-1} = x_{n-1}) = \int_A f(x_n = x \mid x_{n-1})\, dx$$

where $f(x_n \mid x_{n-1})$ = pdf of ~~[illegible]~~ $N(\theta x_{n-1}, \sigma^2)$

Stationarity: if π is a density s.t.

$$\pi(y) = \int K(y \mid x)\, \pi(x)\, dx$$

then π is the stationary density for the M.C. defined by K.

If the current state of the chain is drawn from π, then marginal density of next state is also π. (Analogous to discrete state space M.C.'s).

Irreducibility

M.C. can reach all interesting (positive prob.) regions (sets/intervals) in the state space.

Discrete case: $\forall i, j \in \Omega, \exists n$ s.t. $P_{ij}^{n} > 0$.

Continuous state space: π-irreducibility

Let $\pi(A) = \int_A \pi(x)\, dx$ (slight abuse of notation π)

M.C. is π-irreducible if $\forall x \in \Omega$ and all A s.t. $\pi(A) > 0$, $\exists n$ s.t. $P^n(x, A) > 0$.

That is, any set w/ positive prob. under π is accessible from every pt. in state space.