Metropolis – Hastings Algorithm (M-H)

A general approach for constructing a Markov chain that has the desired stationary distribution ($\pi_j = \pi(j)$)

1) Proposal distribution:

Assume that $X^t = i$. Need to propose a new state with distribution $q_{ij} = q(j|i)$.

2) Calculate the Hastings' ratio

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

3) Acceptance/Reject step

Generate $U \sim U(0,1)$ and set

$$X^{t+1} = \begin{cases} j & \text{if } U \le a_{ij} \\ i\left(= X^t\right) & \text{otherwise} \end{cases}$$

Notes:

1) Gibbs sampling is a special case of M-H as for each step,

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} = 1$$

which implies the relationship also holds for a complete scan through all the variables.

2) The Metropolis (Metropolis et al, 1953) algorithm was based on a symmetric proposal distribution ($q_{ij} = q_{ji}$)

$$\alpha_{ij} = \min\left\{\frac{\pi_j}{\pi_i}, 1\right\}$$

So a higher probability state will always be accepted.

3) As with many other sampling procedures, π and q only need to be known up to normalizing constants as they will be cancelled out when calculating the Hastings' ratio.

4) Periodicity isn't a problem usually.

   For many proposals, $q_{ii} > 0$ for all i. Also if $a_{ij} < 1$, $P\left[X^{t+1} = i \middle| X^t = i\right] > 0$, thus some states have period 1, which implies the chain is aperiodic.

5) $q_{ij}a_{ij}$ gives the 1-step transition probabilities of the chain (e.g. its $p(x|y)$ in the earlier notation).

6) Detailed balance is easy. Without loss of generality, assume that

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} < 1$$

   (which implies $a_{ij} < 1$ and $a_{ji} = 1$)

   Then

$$\begin{aligned} \pi_i q_{ij} a_{ij} &= \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \\ &= \pi_j q_{ji} \\ &= \pi_j q_{ji} a_{ji} \end{aligned}$$

7) The big problem is irreducibility. However by setting the proposal to correspond to a irreducible chain solves this.

Proposal distribution ideas:

1) Approximate the distribution. For example use a normal with similar means and variances. Or use a t with a moderate number of degrees of freedom.

2) Random walk

$$q(y|x) = q(y - x)$$

If there is a continuous state process, you could use

$$y = x + \varepsilon; \quad \varepsilon \sim q(\bullet)$$

For a discrete process, you could use

$$q(j|i) = \begin{cases} 0.4 & j = i - 1 \\ 0.2 & j = i \\ 0.4 & j = i + 1 \end{cases}$$

3) Autoregressive chain

$$y = a + B(x - a) + z; \quad z \sim q(\bullet)$$

For the random walk and autoregressive chains, q does not need to correspond to a symmetric distribution (though that is common).

4) Independence sampler

$$q(y|x) = q(y)$$

For an independence sampler you want q to be similar to π.

$$a_{ij} = \min\left\{\frac{\pi_j q_i}{\pi_i q_j}, 1\right\}$$

If they are too different, q_i/π_i could get very small, making it difficult to move from state i. (The chain mixes slowly).

5) Block at a time

Deal with variables in blocks like the Gibbs sampler. Sometimes referred to Metropolis within Gibbs.

Allows for complex problems to be broken down into simpler ones.

Any M-H style update can be used within each block (e.g. random walk for one block, independence sampler for the next, Gibbs for the one after that).

Allows for a Gibbs style sampler, but without the worry about conjugate distributions in the model to make sampling easier.

Pump Example:

$$s_i \mid \lambda_i \sim \text{Poisson}(\lambda_i t_i)$$

$$\lambda_i \mid \mu, \sigma^2 \sim \text{LogN}(\mu, \sigma^2)$$

$$\mu \sim N(\nu, \tau^2)$$

$$\sigma^2 \sim \text{IGamma}(\gamma, \delta)$$

Can perform Gibbs on μ and σ^2 but not on λ, due the non-conjugacy of the Poisson and log Normal distributions.

Step i, $i = 1, \ldots, 10$ (M-H):

Sample λ_i from $\lambda_i \left| s, \mu, \sigma^2 \right.$ with proposal

$\lambda_i^* \sim \log\mathrm{N}\left(\lambda_i, \theta^2\right)$ (Multiplicative random walk)

$$HR = \frac{\left(\lambda_i^* t_i\right)^{s_i} e^{-\lambda_i^* t_i} \frac{1}{\lambda_i^* \sigma} \phi\left(\frac{\log \lambda_i^* - \mu}{\sigma}\right)}{\left(\lambda_i t_i\right)^{s_i} e^{-\lambda_i t_i} \frac{1}{\lambda_i \sigma} \phi\left(\frac{\log \lambda_i - \mu}{\sigma}\right)} \times \frac{\frac{1}{\lambda_i \theta} \phi\left(\frac{\log \lambda_i - \log \lambda_i^*}{\theta}\right)}{\frac{1}{\lambda_i^* \theta} \phi\left(\frac{\log \lambda_i^* - \log \lambda_i}{\theta}\right)}$$

$$a_{ij} = \min\left(HR, 1\right)$$

Step 11 (Gibbs):

Sample μ from $\mu \left| \lambda, \sigma^2, \nu, \tau^2 \sim N\left(\text{mean}, \text{var}\right)\right.$

where

$$\text{mean} = \text{var}\left(\frac{1}{\sigma^2} \sum \log \lambda_i + \frac{\nu}{\tau^2} \right)$$

$$\text{var} = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}$$

Step 12 (Gibbs):

Sample σ^2 from

$$\sigma^2 \left| \lambda, \mu, \gamma, \delta \right.$$

$$\sim \text{IGamma}\left(\gamma + 5, \delta + \frac{1}{2} \sum \left(\log \lambda_i - \mu \right)^2 \right)$$

Parameters for run

Burn-in: 1000

Imputations: 100,000

ν = -50

τ^2 = 100

γ = 1

δ = 100

θ^2 = 0.01

Starting values

$$\lambda_i = l_i$$

$$\mu = \frac{1}{10}\sum \log l_i$$

$$\sigma^2 = \frac{1}{9}\sum \left(\log l_i - \mu\right)^2$$

Other options

1) Combine steps 1 – 10 into a single draw.

   With this option all λs change or none do. In the sampler used, whether each λ changes is independent of the other λs.

   The option used is probably preferable, as it should lead to better mixing of the chain.

2) Combine sampling λ, μ, and σ^2 into a single M-H step. Probably suboptimal as the proposal distribution won't be a great match for the joint posterior distribution of λ, μ, and σ^2.

Rejection rates

Having some rejection can be good.

With the multiplicative random walk sampler used, if θ^2 is too small, there will be very few rejections, but the sampler will move too slowly through the space.

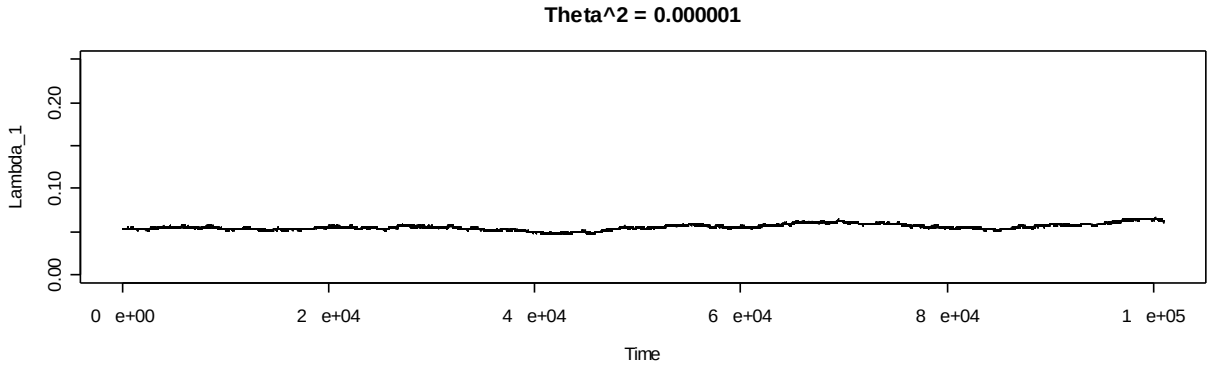
Increasing θ^2 will lead to better mixing, as bigger jumps can be made, though it will lead to higher rejection rates.

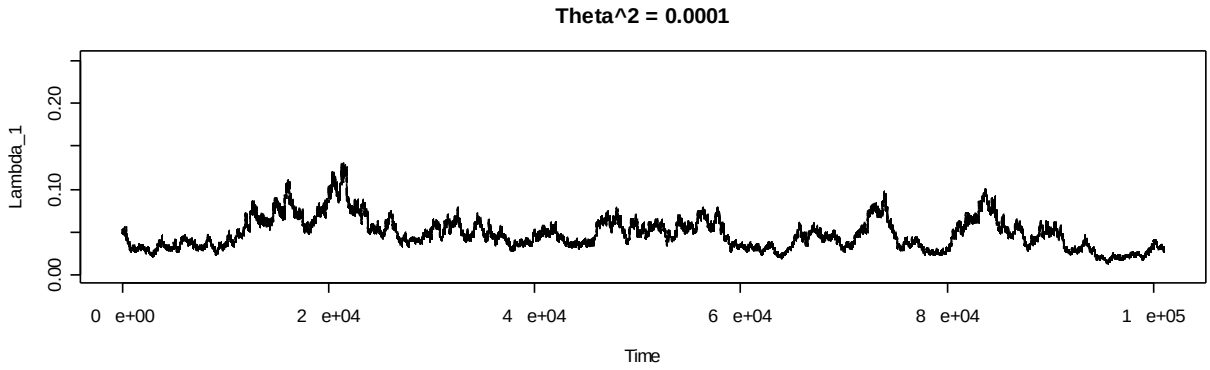
You need to find a balance between rejection rates, mixing of the chain, and coverage of the state space.

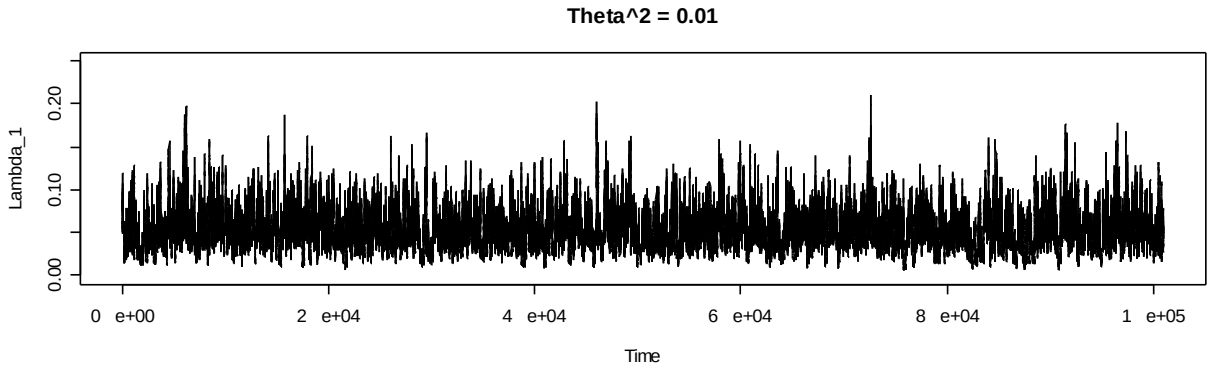
For some problems, a rejection rate of 50% is fine and I've seen reports for large problems using normal random walk proposals the rejection rates of 75% are optimal.

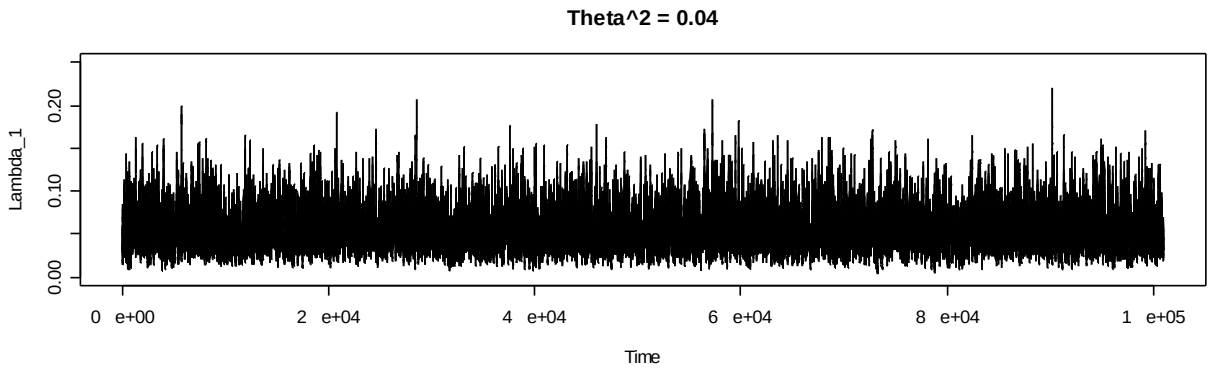
Rejection rates for failure rates proposals under different random walk variances

Pump	0.000001	0.0001	0.01	0.04
1	0.00012	0.00613	0.07045	0.13776
2	0.00009	0.00531	0.03141	0.06130
3	0.00034	0.00784	0.07107	0.13754
4	0.00043	0.01126	0.11705	0.22482
5	0.00028	0.00691	0.05521	0.10705
6	0.00126	0.01442	0.13511	0.26028
7	0.00012	0.00148	0.03027	0.05735
8	0.00007	0.00414	0.02854	0.05824
9	0.00024	0.00559	0.06105	0.12131
10	0.00070	0.01461	0.14790	0.27735


Theta^2 = 0.000001
Lambda_1
0.00
0.10
0.20
0 e+00
2 e+04
4 e+04
6 e+04
8 e+04
1 e+05
Time


Theta^2 = 0.0001
Lambda_1
0.00
0.10
0.20
0 e+00
2 e+04
4 e+04
6 e+04
8 e+04
1 e+05
Time


Theta^2 = 0.01
Lambda_1
0.00
0.10
0.20
0 e+00
2 e+04
4 e+04
6 e+04
8 e+04
1 e+05
Time


Theta^2 = 0.04
Lambda_1
0.00
0.10
0.20
0 e+00
2 e+04
4 e+04
6 e+04
8 e+04
1 e+05
Time

Standard errors in MCMC

As discussed before, the correlation of the chain of the chain must be taken into account when determining standard errors of quantities estimated by the sampler.

Suppose we use $\bar{x}$ to estimate and that the burn-in period was long enough to get into the stationary distribution. Then

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n^2}\left(n + 2\sum_{j=1}^{n-1}(n-j)\,\rho_j \right)$$

For a reasonable chain, the autocorrelations will die off and so lets assume that they will be negligible for $j > K$. Then the above reduces to

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n^2}\left(n + 2\sum_{j=1}^{K}(n-j)\,\rho_j \right)$$

If the autocorrelations die off fairly quickly, σ^2 and ρ_j can be estimated consistently (though with some bias) by the usual empirical moments.

Another approach is blocking. Assume that $n = Jm$ for integers J and m. Then let

$$\tilde{x}_j = \frac{1}{m} \sum_{i=(j-1)m+1}^{jm} x_i; \quad j = 1, \ldots, J$$

Note that $\bar{x} = \bar{\tilde{x}}$. If m is large relative to K, then the correlations between the $\tilde{x}_j$ should negligible and the variance can be estimated as if the $\tilde{x}_j$ were independent.

If the correlation is slightly larger, it might be reasonable to assume that the correlation between $\tilde{x}_j$ and $\tilde{x}_{j+1}$ is some value ρ to be determined, but that correlations at larger lags are negligible. In this case

$$\mathrm{Var}(\bar{x}) \doteq \mathrm{Var}(\tilde{x}_j) \frac{1 + 2\rho}{J}$$

Estimates with $m = 100$

Parameter	$\bar{x}$	SE	ρ
λ_1	0.05290	0.00071	0.36116
λ_2	0.06926	0.00277	0.66197
λ_3	0.07837	0.00106	0.35354
λ_4	0.11053	0.00056	0.10520
λ_5	0.56167	0.01119	0.46975
λ_6	0.60546	0.00237	0.10960
λ_7	0.92318	0.04068	0.67346
λ_8	0.90361	0.03766	0.63510
λ_9	1.82900	0.02884	0.33629
λ_{10}	2.10188	0.00726	0.05263
μ	-2.52492	0.01384	0.41517
σ^2	27.15958	0.09967	0.07579

Estimates with $m = 1000$

Parameter	$\bar{x}$	SE	ρ
λ_1	0.05290	0.00075	0.13239
λ_2	0.06926	0.00399	0.18756
λ_3	0.07837	0.00088	-0.13079
λ_4	0.11053	0.00045	-0.15794
λ_5	0.56167	0.01205	-0.00838
λ_6	0.60546	0.00226	-0.07845
λ_7	0.92318	0.06081	0.12201
λ_8	0.90361	0.04822	0.04495
λ_9	1.82900	0.03303	0.07779
λ_{10}	2.10188	0.00757	0.06487
μ	-2.52492	0.01981	0.15224
σ^2	27.15958	0.13956	0.29726