Metropolis-Hasting Algorithm

Jesús Fernández-Villaverde
University of Pennsylvania

Building our McMc

Our previous chapter showed that we need to find a transition Kernel $P(x, A)$ such that:

1. It is time reversible.

2. It is irreducible.

3. It is aperiodic.

4. (Bonus Points) It is Harris-recurrent and Geometrically Ergodic.

History of Metropolis-Hastings

- Original contribution: Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953).

- Generalized by Hastings (1970).

- Ignored for a long time: Hastings did not get tenure!

- Rediscovered by Tanner and Wong (1987) and Gelfand and Smith (1990).

Metropolis-Hastings Transition Kernel

Let:

$$P_{MH}(x, dy) = p_{MH}(x, y)\, dy + r_{MH}(x)\, \delta_x(dy)$$

where:

$$p_{MH}(x, y) = q(x, y)\, \alpha(x, y)$$

$$\alpha(x, y) = \min\left\{ \frac{f(y)\, q(y, x)}{f(x)\, q(x, y)}, 1 \right\}$$

and $q(x, y)$ is a candidate-generating density that is irreducible and aperiodic.

Lemma

$\int_A f(y)\,dy = \int_{\mathcal{X}} P_{MH}(x, A)\, f(x)\,dx.$

Proof: We only need to show that

$$f(x)\, p_{MH}(x, y) = f(y)\, p_{MH}(y, x)$$

Assume without loss of generality that:

$$\alpha(x, y) < 1 \Rightarrow \alpha(y, x) = 1$$

Then:

$$f(x)\, p_{MH}(x, y) = f(x)\, q(x, y)\, \alpha(x, y)$$

$$= f(x)\, q(x, y) \min\left\{\frac{f(y)\, q(y, x)}{f(x)\, q(x, y)}, 1\right\} = f(x)\, p(x, y)\, \frac{f(y)\, q(y, x)}{f(x)\, q(x, y)}$$

$$= f(y)\, q(y, x) = f(y)\, p_{MH}(y, x)$$

Remark

- Why do we need the min operator?

- In general

$$f(x)\, q(x,y) \neq f(y)\, q(y,x)$$

  If, for example $f(x)\, q(x,y) > f(y)\, q(y,x)$, the process moves from x to y too often and from y to x too rarely.

- We correct this problem with the probability $\alpha(\cdot,\cdot)$.

- Now we have reduced the number of moves from x to y.

Symmetric Candidate-Generating Densities

- We can take a candidate-generating density $q(x,y) = q(x,y)$ (for example a Random walk). Then:

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

- Then, if the jump is "uphill" $(f(y)/f(x) > 1)$, we always accept:

$$\alpha(x,y) = 1 \Rightarrow p_{MH}(x,y) = q(x,y) \Rightarrow r_{MH}(x) = 0$$

- If the jump is "downhill" $(f(y)/f(x) < 1)$, we accept with nonzero probability:

$$\alpha(x,y) < 1 \Rightarrow p_{MH}(x,y) < p(x,y) \Rightarrow r_{MH}(x) > 0$$

Pseudo-Code

1. Initialize the algorithm with an arbitrary value x_0 and M.

2. Set $j = 1$.

3. Generate x_j^* from $q\left(x_{j-1}, x_j^*\right)$ and u from $\mathcal{U}[0,1]$.

4. If $u \leq \alpha\left(x_{j-1}, x_j^*\right)$ then $x_j = x_j^*$, if $u > \alpha\left(x_{j-1}, x_j^*\right)$ then $x_j = x_{j-1}$.

5. If $j \leq M$ then $j \rightsquigarrow j+1$ and got to 3.

Remarks on Metropolis-Hasting

- Metropolis-Hasting Algorithm is defined by $q(x, y)$. Alternatives?
- We need to able to evaluate a function $g(x) \propto f(x)$. Since we only need to compute the ratio $f(y)/f(x)$, the proportionality constant is irrelevant.
- Similarly, we only care about $q(\cdot)$ up to a constant.
- If the candidate is rejected, the current value is taken as the next value in the sequence. Note difference with acceptance sampling.

Choosing $q(x, y)$ I

- A popular choice for $q(x, y)$ is a random walk:

$$y = x + \varepsilon \text{ where } \varepsilon \sim \mathcal{N}(0, \Sigma)$$

- It is known as a *Random-Walk M-H*.
- Random walk satisfies all conditions of a good transition kernel.
- Good default option.
- How do we determine Σ? Hessian of distribution of interest.

Choosing $q(x, y)$ II

- Another popular choice is the *Independent M-H.*
- We just make $q(x, y) = g(y)$.
- Note similarity with acceptance sampling.
- However, the independent M-H accepts more often.
- If $f(x) \leq ag(x)$, then the independent M-H will accept at least $1/a$ of the proposals.

Theoretical Properties

- We built time reversibility on our definition of the M-H transition kernel. What about other requirements?

- If $q(x, y)$ is positive and continuous and $\mathcal{X}$ is connected, we have a Law of Large Numbers.

- If $q(x, y)$ is not reversible (general case), we have an aperiodic chain, which delivers convergence.

- If invariant distribution is bounded, we will also have a Central Limit Theorem.

Example I (`t_metropolis.m`)

- We revisit our problem of drawing from a t distribution.
- Remember how difficult it was to use, for example, a normal distribution to sample as an envelope?
- Basically, because dealing with tails with difficult.
- Now, we will see that with a Metropolis-Hastings the problem is quite simple.

Example I

- Use MH to provide a numerical approximation to $t\,(x, 3)$, a t distribution with 3 degrees of freedom, evaluated at x.

- We need to get a random draw $\left\{x_j\right\}_{j=1}^{N}$ from $t(3)$ using the MH.

- Implemented in my code `t_metropolis.m`.

Pseudocode

1. Initialize the algorithm with $x_0 = 0$ and M. Set $j = 1$.

2. Generate $x_j^* = x_{j-1} + \mathcal{N}(0, 3.4)$.

3. Then $\alpha\left(x_{j-1}, x_j^*\right) = \min\left\{f_3\left(x_j^*\right) / f_3\left(x_{j-1}\right), 1\right\}$.

4. Generate $u \sim \mathcal{U}[0, 1]$.

5. If $u \leq \alpha\left(x_{j-1}, x_j^*\right)$ then $x_j = x_j^*$, otherwise $x_j = x_{j-1}$.

6. If $j \leq M$ then $j \rightsquigarrow j + 1$ and go to 3.

Output

- Output is a random draw $\left\{x_j\right\}_{j=1}^{M}$.

- With simulation, we can compute CDF:

$$\Phi(t) \simeq \frac{1}{M} \sum_{i=1}^{M} \delta_{\{x_i : x_i < t\}}(x_i)$$

- Some times, researchers report a smoothed version of the density (for example with a Kernel estimator).

- Similarly, we can compute the integral of any function of interest and Numerical errors.

Rate of Convergence

At which speed does the Chain converge? How long the Chain should run? Three important things to do:

- Run a set of different Chains with different initial values and compare within and between Chains variation.
- Check serial correlation of the draws.
- Make M an increasing function of the serial correlation of the draws.
- Run N different chains of length M with random initial values and take the last value of each chain.

Diagnosis of Convergence

- Can we check convergence of our estimates?
- Different procedures. See Robert and Casella (2004) chapter 12.
- My experience with formal convergence criteria is not a happy one: they accept convergence too quickly.
- Graphical alternatives: time series and recursive means.

More on Convergence

- Often, convergence takes longer than what you think.
- Case of bi-modal distributions.
- Play it safe: just let the computer run a few more times.
- Use acceleration methos or Rao-Blacwellization

$$var\left(E\left(h\left(X|Y\right)\right)\right) \leq varh\left(X\right)$$

.

One Chain versus Many Chains

- Should we use one long chain or many different chains?
- The answer is clear: only one long chain.
- Fortunately, the old approach of many short chains is disappearing.
- This does not mean that you should not do many runs while you are tuning your software!

Burn-in

- Should we burn-in the first simulations?
- Common practice.
- However, link with theory is tenuous at best.
- Just a way to determine an initial value of the parameter.
- There are better ways to proceed.

Acceptance Ratio

- If the candidate is rejected, the current value is taken as the next value in the sequence.

- Choosing the acceptance ratio is important for a good numerical performance of the algorithm.

- Acceptance rate should be (Roberts, Gelman and Gilks, 1994):

  1. 45% for unidimensional problems.

  2. 23% in the limit.

  3. 25% for 6 dimensions.

Example II (`metropolis.m`)

- Simulate 200 observations from the following model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t = w_t' \Phi + \epsilon_t$$

where $\phi_1 = 1$, $\phi_2 = -0.5$, $\epsilon_t \sim \mathcal{N}(0, 1)$, $w_t = (y_{t-1}, y_{t-2})'$ and $\Phi = (\phi_1, \phi_2)'$.

- Let $S = \{\Phi : \phi_1 + \phi_2 < 1, -\phi_1 + \phi_2 < 1, \phi_2 > -1\}$ define the stationary restrictions of the AR(2) process.

Example II

- Write the likelihood function of $Y = (y_3, y_4, ..., y_{100})$ conditional on y_1 and y_2.

$$\ell(Y|\Phi, \sigma, y_2, y_1) = \left(\sigma^2\right)^{-49} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=3}^{100}\left(y_t - w_t'\Phi\right)\right)$$

- Priors are:
  - $\Phi \in S$
  - $\sigma \in T = \{\sigma : \sigma > 0\}$

Example II

- Posterior:

$$\pi\left(\Phi,\sigma|Y,y_2,y_1\right)\propto\left(\sigma^2\right)^{-49}\exp\left(-\frac{1}{2\sigma^2}\sum_{t=3}^{100}\left(y_t-w_t'\Phi\right)\right)\Upsilon_S\left(\Phi\right)\Upsilon_T\left(\sigma\right)$$

- Let

$$p_{MH}\left(\Phi,\sigma,\Phi',\sigma'\right)=\begin{cases}\phi_1'=\phi_1+\mathcal{U}\left[-0.04,0.04\right]\\ \phi_2'=\phi_2+\mathcal{U}\left[-0.04,0.04\right]\\ \sigma=\sigma+\mathcal{U}\left[-0.004,0.004\right]\end{cases}$$

Pseudocode I

1. Initialize Φ_0, σ_0, and M to some fixed value and set $j = 1$.

2. Set

$$\phi^*_{i,j} = \phi_{i,j-1} + \mathcal{U}\left[-0.04, 0.04\right]$$

for $= 1, 2$ and

$$\sigma^*_j = \sigma_{j-1} + \mathcal{U}\left[-0.004, 0.004\right]$$

3. Then

$$\alpha\left(\Phi_{j-1}, \sigma_{j-1}, \Phi^*_j, \sigma^*_j\right) = \min\left\{\frac{\pi\left(\Phi^*_j, \sigma^*_j | Y, y_2, y_1\right)}{\pi\left(\Phi_{j-1}, \sigma_{j-1} | Y, y_2, y_1\right)}, 1\right\}$$

Pseudocode II

4. Generate $u \sim \mathcal{U}[0,1]$.

5. If $u \leq \alpha\left(\Phi_{j-1}, \sigma_{j-1}, \Phi_j^*, \sigma_j^*\right)$ then $\Phi_j = \Phi_j^*$ and $\sigma_j = \sigma_j^*$, if $u > \alpha\left(\Phi_{j-1}, \sigma_{j-1}, \Phi_j^*, \sigma_j^*\right)$ then $\Phi_j = \Phi_{j-1}$ and $\sigma_j = \sigma_{j-1}$.

6. If $j \leq M$ then $j \rightsquigarrow j+1$ and got to 3.

Example III (`fisher_mcmc.m`)

- Our first real example of how to estimate a DSGE model.
- Model with investment-specific technological change.
- Based on Greenwood, Hercowitz, and Krusell (1997) and Fisher (2004).
- Why investment-specific technological change?

Model I

- Representative household.

- Preferences:

$$E_0 \sum_{t=0}^{\infty} \beta^t \left(\log C_t + \psi \log((1 - L_t)) \right.$$

- Resource constraint:

$$C_t + X_t = A_t K_t^{\alpha} L_t^{1-\alpha}$$

- Law of motion for capital:

$$K_{t+1} = (1 - \delta) K_t + V_t X_t,$$

Model II

- Technology evolution:

$$\begin{aligned} A_t &= e^{\gamma+\varepsilon_{at}} A_{t-1},\ \gamma \geq 0 \text{ and } \varepsilon_{at} \sim \mathcal{N}(0, \sigma_a) \\ V_t &= e^{\upsilon+\varepsilon_{vt}} V_{t-1},\ \upsilon \geq 0 \text{ and } \varepsilon_{vt} \sim \mathcal{N}(0, \sigma_v) \end{aligned}$$

- Note:

  1. Two unit roots.

  2. Different drifts.

  3. Cointegration relations among nominal variables but not among real ones.

Model III

- Technology evolution:

$$\begin{aligned} A_t &= e^{\gamma+\varepsilon_{at}} A_{t-1},\ \gamma \geq 0 \text{ and } \varepsilon_{at} \sim \mathcal{N}(0, \sigma_a) \\ V_t &= e^{\upsilon+\varepsilon_{vt}} V_{t-1},\ \upsilon \geq 0 \text{ and } \varepsilon_{vt} \sim \mathcal{N}(0, \sigma_\upsilon) \end{aligned}$$

- Note:

  1. Two unit roots.

  2. Different drifts.

  3. Cointegration relations among nominal variables but not among real ones.

Transforming Model I

- The previous model is nonstationary because of the presence of two unit roots.

- We need to transform the model into a stationary problem.

- We select a predetermined scaling variable that is fully known before the current period shocks are realized.

Transforming Model II

- We begin with the resource constraint and the law of motion for capital:

$$C_t + X_t = A_t K_t^{\alpha} L_t^{1-\alpha}$$
$$\frac{K_{t+1}}{V_t} = (1-\delta)\frac{K_t}{V_t} + X_t.$$

- If we divide both equations by $Z_t = A_{t-1}^{\frac{1}{1-\alpha}} V_{t-1}^{\frac{\alpha}{1-\alpha}} = \left(A_{t-1} V_{t-1}^{\alpha}\right)^{\frac{1}{1-\alpha}}$, we find:

$$\begin{aligned}
\frac{C_t}{Z_t} + \frac{X_t}{Z_t} &= \frac{A_t V_{t-1}^{\alpha}}{Z_t^{1-\alpha}} \left(\frac{K_t}{Z_t V_{t-1}}\right)^{\alpha} L_t^{1-\alpha} \\
\frac{K_{t+1}}{Z_{t+1} V_t}\frac{Z_{t+1}}{Z_t} &= (1-\delta)\frac{K_t}{Z_t V_{t-1}}\frac{V_{t-1}}{V_t} + \frac{X_t}{Z_t}.
\end{aligned}$$

Transforming Model III

First, note that since:

$$Z_{t+1} = A_t^{\frac{1}{1-\alpha}} V_t^{\frac{\alpha}{1-\alpha}} = A_{t-1}^{\frac{1}{1-\alpha}} V_{t-1}^{\frac{\alpha}{1-\alpha}} e^{\frac{\gamma+\alpha\upsilon+\varepsilon_{at}+\alpha\varepsilon_{\upsilon t}}{1-\alpha}},$$

we have that:

$$\frac{Z_{t+1}}{Z_t} = \frac{A_{t-1}^{\frac{1}{1-\alpha}} V_{t-1}^{\frac{\alpha}{1-\alpha}} e^{\frac{\gamma+\alpha\upsilon+\varepsilon_{at}+\alpha\varepsilon_{\upsilon t}}{1-\alpha}}}{A_{t-1}^{\frac{1}{1-\alpha}} V_{t-1}^{\frac{\alpha}{1-\alpha}}} = e^{\frac{\gamma+\alpha\upsilon+\varepsilon_{at}+\alpha\varepsilon_{\upsilon t}}{1-\alpha}}.$$

Also $\frac{A_t V_{t-1}^{\alpha}}{Z_t^{1-\alpha}} = e^{\gamma+\varepsilon_{at}}$, $\frac{V_{t-1}}{V_t} = e^{-\upsilon-\varepsilon_{\upsilon t}}$, and $Z_t V_{t-1} = A_{t-1}^{\frac{1}{1-\alpha}} V_{t-1}^{\frac{1}{1-\alpha}}$.

Transforming Model IV

Define $\widetilde{C}_t = \frac{C_t}{Z_t}$, $\widetilde{X}_t = \frac{X_t}{Z_t}$ and $\widetilde{K}_t = \frac{K_t}{Z_t V_{t-1}}$. Then:

$$\widetilde{C}_t + \widetilde{X}_t = e^{\gamma+\varepsilon_{at}} \widetilde{K}_t^{\alpha} L_t^{1-\alpha}$$

$$e^{\frac{\gamma+\alpha\upsilon+\varepsilon_{at}+\alpha\varepsilon_{\upsilon t}}{1-\alpha}} \widetilde{K}_{t+1} = (1-\delta)\, e^{-\upsilon-\varepsilon_{\upsilon t}} \widetilde{K}_t + \widetilde{X}_t$$

or, summing both expressions:

$$\widetilde{C}_t + e^{\frac{\gamma+\alpha\upsilon+\varepsilon_{at}+\alpha\varepsilon_{\upsilon t}}{1-\alpha}} \widetilde{K}_{t+1} = e^{\gamma+\varepsilon_{at}} \widetilde{K}_t^{\alpha} L_t^{1-\alpha} + (1-\delta)\, e^{-\upsilon-\varepsilon_{\upsilon t}} \widetilde{K}_t$$

New Model

$$E_0 \sum_{t=0}^{\infty} \beta^t \left(\log \widetilde{C}_t + \psi \log(1 - L_t) \right).$$

such that

$$\widetilde{C}_t + e^{\frac{\gamma + \alpha\upsilon + \varepsilon_{at} + \alpha\varepsilon_{\upsilon t}}{1-\alpha}} \widetilde{K}_{t+1} = e^{\gamma + \varepsilon_{at}} \widetilde{K}_t^{\alpha} L_t^{1-\alpha} + (1 - \delta) e^{-\upsilon - \varepsilon_{\upsilon t}} \widetilde{K}_t$$

with first order conditions:

$$\begin{aligned} \frac{e^{\frac{\gamma + \alpha\upsilon + \varepsilon_{at} + \alpha\varepsilon_{\upsilon t}}{1-\alpha}}}{\widetilde{C}_t} &= \beta E_t \frac{1}{\widetilde{C}_{t+1}} \left(\alpha e^{\gamma + \varepsilon_{at+1}} \widetilde{K}_{t+1}^{\alpha} L_{t+1}^{1-\alpha} + (1 - \delta) e^{-\upsilon - \varepsilon_{\upsilon t+1}} \right) \\ \psi \frac{\widetilde{C}_t}{1 - L_t} &= (1 - \alpha) e^{\gamma + \varepsilon_{at}} \widetilde{K}_t^{\alpha} L_t^{-\alpha} \end{aligned}$$

Observables

- Certain degree of arbitrariness.
- Use of educated economic intuition.
- First differences of output and hours.

Four Remarks

- Other transformation methods? Prefiltering. Hansen and Sargent (1993).

- Relation with stochastic singularity. Measurement errors. Advantages and disadvantages of measurement errors.

- Initialization of the filter. Alternatives to first differences.

- What happened with our cointegration relations?

Multiple-block Metropolis-Hastings

- Often we can group different variables of the vector $x \sim f(x)$ in one block.

- Why? Increases efficiency, especially in large spaces.

- Furthermore, we will see that the Gibbs Sampler is a case of Multiple-block Metropolis-Hastings.

Partitions

- Partition x into $x = \{x_1, ..., x_p\}$.
- Define: x_{-k} to be all the blocks excluding x_k.
- Define $f\left(x_k, x_{-k}\right)$ to be the joint density of x regardless of where x_k appears in the list of blocks.
- Define $\{q_k\left(x_k, y_k | x_{-k}\right), k \leq p\}$ to be a family of candidate-generating densities.

Transition Densities

- Define

$$\alpha_k\left(x_k, y_k | x_{-k}\right) = \min\left\{\frac{f\left(y_k, x_{-k}\right) q_k\left(y_k, x_k | x_{-k}\right)}{f\left(x_k, x_{-k}\right) q_k\left(x_k, y_k | x_{-k}\right)}, 1\right\}$$

  as the probability of moving within one block.

- The algorithm is a M-H where we do the update block by block.

- The update of x_k is done in each step of the cycle.

- Why does it work? Local time reversibility.

Monte Carlo Optimization

- We saw that, at the core of the M-H, there is an "up-hill" climber: we always accept if we go up.
- But there is a great twist: sometimes we go down.
- Can we use this idea for optimization?
- Is there a theory behind this?
- Yes, Monte Carlo Optimization.

A Motivation Example

- Imagine we have the function:

$$\begin{aligned} h(x,y) &= -\left(x\sin(20y)+y\sin(20x)\right)^2\cosh\left(\sin(10x)\,x\right) \\ &\quad -\left(x\cos(10y)-y\sin(10x)\right)^2\cosh\left(\cos(20y)\,y\right) \end{aligned}$$

- Let's take a look: `crazy_function.m`

- Do you think we can maximize this function using a Newton-type algorithm?

- Maybe with a stochastic gradient method. But, does not a stochastic gradient motivate a different approach?

Simulated Annealing

- Developed by Kirkpatrick, Gellat, and Vecchi (1983).
- They built on Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953)!
- Name comes from Metallurgy. By slowly cooling down an anneal, we get a configuration of atoms with the lowest internal energy (the kinetic energy of the molecules.
- Natural counterpart in optimization problems.

Main Idea

- Given a parameter $T > 0$ (often called *temperature*), we draw x^*_{j+1} from:

$$\exp\left(f\left(x^*_j\right)/T\right)$$

- Function f is the function to maximize.

- As T goes to zero, the values simulated from this distribution become more concentrated around a narrower neighborhood of the local maxima of f.

Pseudocode

1. Initialize the algorithm with an arbitrary value x_0 and M.

2. Set $j = 1$.

3. Generate x_j^* from symmetric $q\left(x_{j-1}, x_j^*\right)$ and u from $\mathcal{U}[0,1]$.

4. If $u \leq \left\{\exp\left(\Delta f\left(x_j^*\right)/T_j\right), 1\right\}$ then $x_j = x_j^*$, otherwise $x_j = x_{j-1}$.

5. If $j \leq M$ then $j \rightsquigarrow j+1$, $T_j \rightsquigarrow T_{j+1}$, and go to 3.

Intuition

- If we improve, we always move into the proposed direction.
- However, if we do not improve, we may be tempted to move anyway.
- Why? Because it avoids getting stuck in a local maxima.
- Example from nature: bees.

Remarks

- Proof of why simulated annealing works is rather involved technically. More general case: Andrieu and Doucet (2001).

- Main practical issue: how to select the sequence of temperatures to go to zero at the right rate.

- Literature suggests rates of $1/\log t$.

Equivalence with Metropolis-Hastings

- From the pseudocode, we see a strong resemblance with M-H.
- Basically Simulated Annealing is a M-H with stationary distribution $\exp\left(f\left(x_j^*\right)/T\right)$ for a fixed T.
- For non-fixed, we need to work a bit more (we will be handling a time-heterogeneous Markov Chain), but the idea is the same
- Think about Harris Recurrence!

Practical Implementation I

- Use your M-H to draw from the posterior as you will otherwise do.
- Keep track of the highest value of the likelihood you have found so far and the parameters that generated it.
- As the number of M-H simulations goes to infinity, you will pass a.s. through the max of the likelihood.
- Simpler to see with uniform priors.

Practical Implementation II

- In practice this means you can do your ML and your Bayesian estimation simultaneously.
- You can feed your max as the initial guess of a Newton-type algorithm
- Relation with the paper of Chernozhukov and Hong (2003) that we have already seen.

Genetic Algorithms I

- Before we used the example of bees.
- Nature seems to be good at optimizing.
- Idea behind evolution.
- Can we apply those insights? Genetic Algorithms.

Genetic Algorithms II

- Developed by John Holland.
- Basic idea: genetic information is copied subject to mutations and there is a survival of the fittest.
- Long tradition in economics: Malthus and Darwin.
- Example: traveling salesman problem.

Remarks

- Solution implies "Punctuated Equilibria".

- Something like this is observed in evolutionary landscapes (Stephen Jay Gould).

- Implications for economics:

  1. Technological change.

  2. Learning environments.