Markov Chain Monte Carlo

Ed Herbst

March 16, 2023

The Metropolis-Hastings Algorithm

- Metropolis-Hastings (MH) algorithm belongs to the class of Markov chain Monte Carlo (MCMC) algorithms.
- Algorithm constructs a Markov chain such that the stationary distribution associated with this Markov chain is unique and equals the posterior distribution of interest.
- First version constructed by Metropolis et al. (1953). Later generalized by Hastings (1970). Tierney (1994) proved important convergence results for MCMC algorithms.
- Introduction: Chib and Greenberg (1995). Textbook Robert and Casella (2004) or Geweke (2005).

Markov Chain Monte Carlo

- Importance sampler generates a sequence of independent draws from the posterior distribution $\pi(\theta)$, the MH algorithm generates a sequence of serially correlated draws.

- As long as the correlation in the Markov chain is not too strong, Monte Carlo averages of these draws can accurately approximate posterior means of $h(\theta)$.

- We are going to care a lot about this correlation. Why?

$$\sqrt{n}(\bar{X} - \mathbb{E}[\bar{X}]) \Longrightarrow N\left(0, \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}[X_i] + \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} COV(X_i, X_j)\right)$$

The Metropolis Hastings Algorithm

A key ingredient is the proposal distribution $q(\vartheta|\theta^{i-1})$, which potentially depends on the draw θ^{i-1} in iteration $i-1$ of the algorithm.

Algorithm (Generic MH Algorithm)

For $i = 1$ to N: Draw ϑ from a density $q(\vartheta|\theta^{i-1})$. Set $\theta^i = \vartheta$ with probability

$$\alpha(\vartheta|\theta^{i-1}) = \min\left\{1,\ \frac{p(Y|\vartheta)p(\vartheta)/q(\vartheta|\theta^{i-1})}{p(Y|\theta^{i-1})p(\theta^{i-1})/q(\theta^{i-1}|\vartheta)}\right\}$$

and $\theta^i = \theta^{i-1}$ otherwise.

Because $p(\theta|Y) \propto p(Y|\theta)p(\theta)$ we can replace the posterior densities in the calculation of the acceptance probabilities $\alpha(\vartheta|\theta^{i-1})$ This yields a Markov transition kernel $K(\theta|\tilde{\theta})$, where the conditioning value $\tilde{\theta}$ corresponds to the parameter draw from iteration $i-1$.

Convergence

Probability theory for MH is much harder than for IS.

- Suppose that $\theta^0 \sim g(\cdot)$ and θ^N is obtained by iterating the Markov transition kernel forward N times, then is it true that θ^N is approximately distributed according to $p(\theta|Y)$ and the approximation error vanishes as $N \longrightarrow \infty$?
- Suppose that (i) is true, is it also true that sample averages of θ^i, $i = 1, \ldots, N$ satisfy a SLLN and a CLT?

Key property: invariance of Markov Chain.

$$p(\theta|Y) = \int K(\theta|\tilde{\theta})p(\tilde{\theta}|Y)d\tilde{\theta}. \quad (1)$$

Show this property using reversibility of the Markov Chain

Not sufficient for SLLN or CLT, these things depend on q and π.

Look at specific example.

A Specific Example

- Suppose the parameter space is discrete and θ can only take two values: τ_1 and τ_2.
- The posterior distribution then simplifies to two probabilities which we denote as $\pi_l = \mathbb{P}\{\theta = \tau_l | Y\}$, $l = 1, 2$.
- The proposal distribution in Algorithm~1 can be represented as a two-stage Markov process with transition matrix

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}, \tag{2}$$

  where q_{lk} is the probability of drawing $\vartheta = \tau_k$ conditional on $\theta^{i-1} = \tau_l$.
- Assume that

$$q_{11} = q_{22} = q, \quad q_{12} = q_{21} = 1 - q$$

  and that the posterior distribution has the property

$$\pi_2 > \pi_1.$$

Deriving the Transition Kernel

- Suppose that $\theta^{i-1} = \tau_1$. Then with probability q, $\vartheta = \tau_1$. The probability that this draw will be accepted is

$$\alpha(\tau_1|\tau_1) = \min\left\{1, \frac{\pi_1/q}{\pi_1/q}\right\} = 1.$$

- With probability $1-q$ the proposed draw is $\vartheta = \tau_2$. The probability that this draw will be rejected is

$$1 - \alpha(\tau_2|\tau_1) = 1 - \min\left\{1, \frac{\pi_2/(1-q)}{\pi_1/(1-q)}\right\} = 0$$

  because we previously assumed that $\pi_2 > \pi_1$.
- The probability of a transition from $\theta^{i-1} = \tau_1$ to $\theta^i = \tau_1$ is

$$k_{11} = q \cdot 1 + (1-q) \cdot 0 = q.$$

Transition Kernel, Continued

- Similar reasoning as before

$$K = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \begin{bmatrix} q & (1-q) \\ (1-q)\frac{\pi_1}{\pi_2} & q + (1-q)\left(1 - \frac{\pi_1}{\pi_2}\right) \end{bmatrix}.$$

- K has two eigenvalues λ_1 and λ_2:

$$\lambda_1(K) = 1, \quad \lambda_2(K) = q - (1-q)\frac{\pi_1}{1-\pi_1}. \tag{3}$$

Eigenvector associated with with $\lambda_1(K)$ determines the invariant distribution of the Markov chain (=posterior). If $\lambda_2(K) \neq 1$, this distribution is unique.

The persistence of the Markov chain is characterized by the eigenvalue $\lambda_2(K)$.

Markov Chain

We can represent the Markov Chain generated by MH as an AR(1). Define:

$$\xi^i = \frac{\theta^i - \tau_1}{\tau_2 - \tau_1}, \quad \xi^i \in \{0, 1\}.$$

ξ^i follows the first-order autoregressive process

$$\xi^i = (1 - k_{11}) + \lambda_2(K)\xi^{i-1} + \nu^i. \tag{4}$$

Conditional on $\xi^{i-1} = j - 1$, $j = 1, 2$, the innovation ν^i has support on k_{jj} and $(1 - k_{jj})$, its conditional mean is equal to zero, and its conditional variance is equal to $k_{jj}(1 - k_{jj})$.

More on Markov Chain

- Persistence of the Markov chain depends on the proposal distribution, which in our discrete example is characterized by the probability q.

- You could get an *iid* sample from the posterior by setting $q = \pi_1$, so $\lambda_2(K) = 0$.)

- OTOH, if $q = 1$, then $\theta^i = \theta^1$ for all i and the equilibrium distribution of the chain is no longer unique.

- General goal of MCMC: keep the persistence of the chain as low as possible.

$$\bar{h}_N = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i)$$

we deduce from a central limit theorem for dependent random variables that

$$\sqrt{N}(\bar{h}_N - \mathbb{E}_\pi[h]) \Longrightarrow N\big(0, \Omega(h)\big),$$

where $\Omega(h)$ is now the long-run covariance matrix

$$\Omega(h) = \lim_{L \longrightarrow \infty} \mathbb{V}_\pi[h]\left(1 + 2\sum_{l=1}^{L}\frac{L-l}{L}\left(q - (1-q)\frac{\pi_1}{1-\pi_1}\right)^l\right).$$

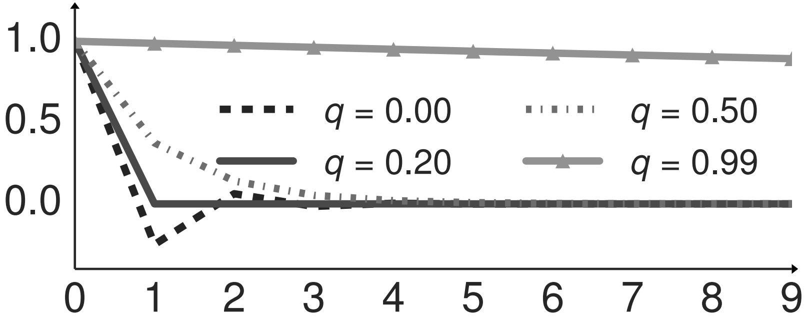
In turn, the asymptotic inefficiency factor is given by

$$\begin{aligned}\mathrm{InEff}_\infty &= \frac{\Omega(h)}{\mathbb{V}_\pi[h]} \qquad (5)\\ &= 1 + 2\lim_{L \longrightarrow \infty}\sum_{l=1}^{L}\frac{L-l}{L}\left(q - (1-q)\frac{\pi_1}{1-\pi_1}\right)^l.\end{aligned}$$

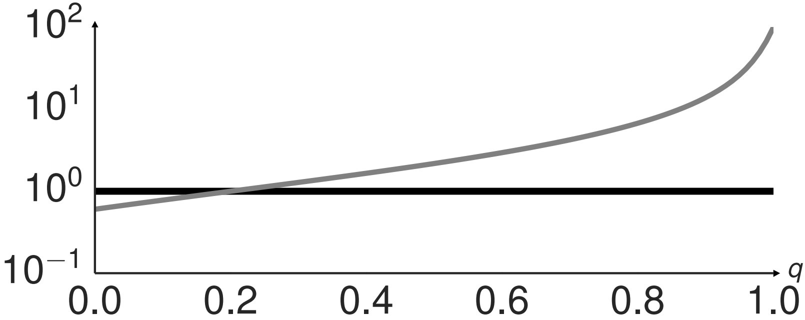
Numerical Example

- Bernoulli distribution $(\tau_1 = 0, \tau_2 = 1)$ with $\pi_1 = 0.2$.
- Assess the effectiveness of different MH settings, we vary $q \in [0, 1)$.
- Look at autocorrelation for $q = \{0, 0.2, 0.5, 0.99\}$.
- Ineff_∞ for $q \in [0, 1)$.
- Relationship between across chain variance and within chain (HAC) estimates. This the heart of many convergence statistics.

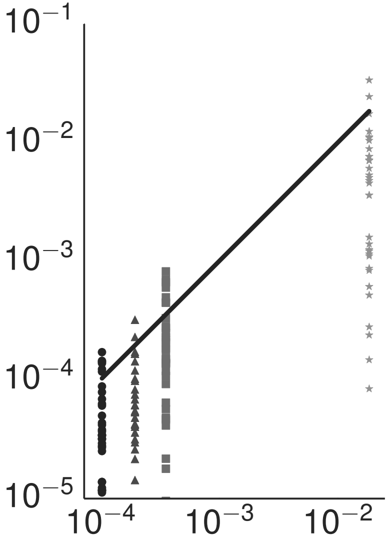
Autocorrelation Functions



Log Inefficiency Factor as function of q



Convergence: within vs across chain variance estimates



Take Aways

- high autocorrelation reflects the fact that it will take a high number of draws to accurately reflect the target distribution
- for large values of q, the variance of Monte Carlo estimates of h drawn from the MH chain are much larger than the variance of estimates derived from *iid* draws
- HAC estimates bracket small-sample estimates, indicating convergence, but they tend to underestimate variance for all q.

How to pick q for a DSGE model?

Random Walk Metropolis-Hastings

- Most popular q for DSGE Models.
- $q(\vartheta|\theta^{i-1})$ can be expressed as the random walk $\vartheta = \theta^{i-1} + \eta$
- η is normally distributed with mean zero and variance $c^2\hat{\Sigma}$.
- Given the symmetric nature of the proposal distribution, the acceptance probability becomes

$$\alpha = \min\left\{\frac{p(\vartheta|Y)}{p(\theta^{i-1}|Y)}, 1\right\}.$$

- Still need to specify c and $\hat{\Sigma}$.

On $\hat{\Sigma}$

- Want $\hat{\Sigma}$ to incorporate information about the posterior.
- One approach: Schorfheide (2000), is to set $\hat{\Sigma}$ to be the negative of the inverse Hessian at the mode of the log posterior, $\hat{\theta}$, obtained by running a numerical optimization .

  This has appealing large sample properties, but can be tedious and innacurate.
- Another (adaptive) approach: use prior variance for a first sequence of posterior draws, the compute the sample covariance matrix and use that as $\hat{\Sigma}$. *Must be fixed eventually*.
- Here we cheat:

$$\text{RWMH-V} : \hat{\Sigma} = \mathbb{V}_\pi[\theta].$$

Picking Scaling c

- Goldilocks principal: choose c so that you don't reject too much or too little.
- Roberts et al. (1997) have derived a limit (in the size of parameter vector) optimal acceptance rate of 0.234 for a special case (normal posterior).
- Most practitioners target an acceptance rate between 0.20 and 0.40.
- Requites pre-estimation tuning.

From Prior to Posterior

- Prior distributions are used to describe the state of knowledge about the parameter vector θ before observing the sample Y.
- In our example, we have to specify a joint probability distribution in 13-dimensional parameter space.

Eliciting prior distributions Del Negro and Schorfheide (2008):

- Group parameters by categories: $\theta_{(ss)}$ (related to steady state), $\theta_{(exo)}$ (related to exogenous processes), $\theta_{(endo)}$ (affects mechanisms but not steady state).

$$
\begin{aligned}
\theta_{(ss)} &= [r^{(A)}, \pi^{(A)}, \gamma^{(Q)}]' \\
\theta_{(exo)} &= [\rho_g, \rho_z, \sigma_g, \sigma_z, \sigma_R]' \\
\theta_{(endo)} &= [\tau, \kappa, \psi_1, \psi_2, \rho_R]'
\end{aligned}
$$

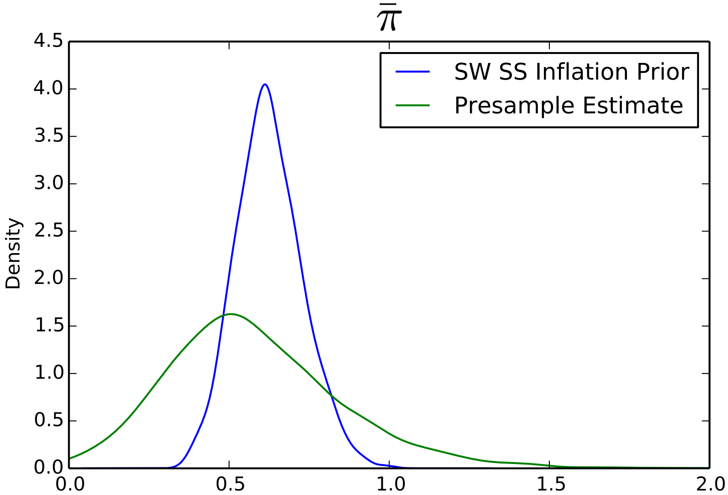
Priors, Continued

- Priors for $\theta_{(ss)}$ are often based on pre-sample averages. If sample starts in 1983:I, the prior distribution for $r^{(A)}$, $\pi^{(A)}$, and $\gamma^{(Q)}$ may be informed by data from the 1970s.
- Priors for $\theta_{(endo)}$ may be partly based on microeconometric evidence.
- Priors for $\theta_{(exo)}$ are the most difficult to specify. You could specific indirectly, by looking at the volatility/autocorrelation of observables implied by $\theta_{(exo)}$ given other parameters.

Above all: Generate draws from the prior distribution of θ; compute important transformations of θ such as steady-state ratios and possibly impulse-response functions or variance decompositions.

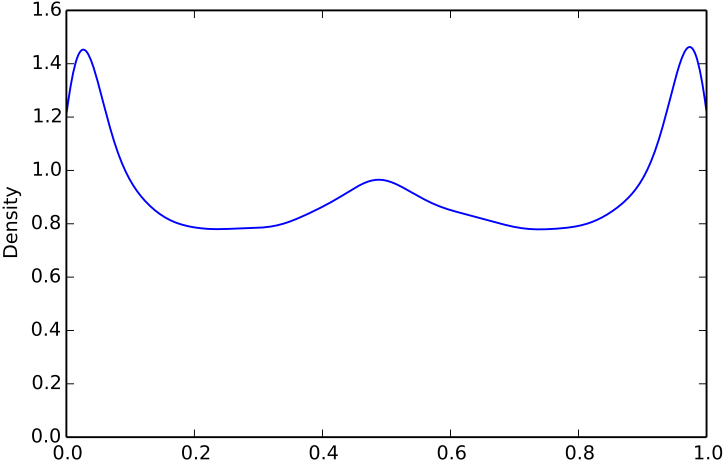
- Marginals may be plausible, while joint is not.
- Nonlinear transformations of uniform variables are not uniform!

Try not to set priors based Y



$$\rho = \frac{x^2}{x^2+y^2}, x \sim U[0,1], y \sim U[0,1]\}$$

Density of ρ



Bayesian Estimation – Prior

Name	Domain	Prior		
		Density	Para (1)	Para (2)
Steady State Related Parameters $\theta_{(ss)}$				
$r^{(A)}$	$\mathbb{R}^+$	Gamma	0.50	0.50
$\pi^{(A)}$	$\mathbb{R}^+$	Gamma	7.00	2.00
$\gamma^{(Q)}$	$\mathbb{R}$	Normal	0.40	0.20
Endogenous Propagation Parameters $\theta_{(endo)}$				
τ	$\mathbb{R}^+$	Gamma	2.00	0.50
κ	$[0, 1]$	Uniform	0.00	1.00
ψ_1	$\mathbb{R}^+$	Gamma	1.50	0.25
ψ_2	$\mathbb{R}^+$	Gamma	0.50	0.25
ρ_R	$[0, 1)$	Uniform	0.00	1.00
Exogenous Shock Parameters $\theta_{(exo)}$				
ρ_G	$[0, 1)$	Uniform	0.00	1.00
ρ_Z	$[0, 1)$	Uniform	0.00	1.00
$100\sigma_R$	$\mathbb{R}^+$	InvGamma	0.40	4.00
$100\sigma_G$	$\mathbb{R}^+$	InvGamma	1.00	4.00
$100\sigma_Z$	$\mathbb{R}^+$	InvGamma	0.50	4.00

Baseline Estimation

Table: Posterior Estimates of DSGE Model Parameters

	Mean	[0.05, 0.95]		Mean	[0.05,0.95]
τ	2.83	[1.95, 3.82]	ρ_r	0.77	[0.71, 0.82]
κ	0.78	[0.51, 0.98]	ρ_g	0.98	[0.96, 1.00]
ψ_1	1.80	[1.43, 2.20]	ρ_z	0.88	[0.84, 0.92]
ψ_2	0.63	[0.23, 1.21]	σ_r	0.22	[0.18, 0.26]
$r^{(A)}$	0.42	[0.04, 0.95]	σ_g	0.71	[0.61, 0.84]
$\pi^{(A)}$	3.30	[2.78, 3.80]	σ_z	0.31	[0.26, 0.36]
$\gamma^{(Q)}$	0.52	[0.28, 0.74]			

Notes: We generated $N = 100,000$ draws from the posterior and discarded the first 50,000 draws. Based on the remaining draws we approximated the posterior mean and the 5th and 95th percentiles.

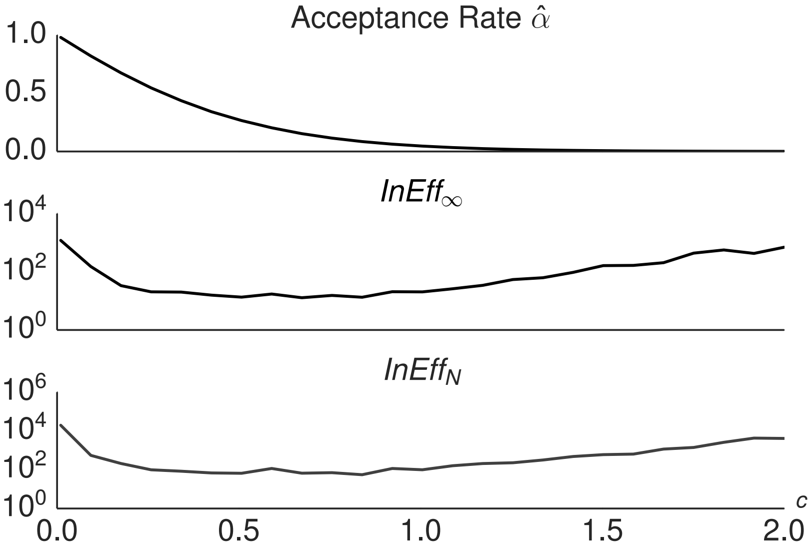
More on c

Vary $c \in (0, 2]$. Look at effect on

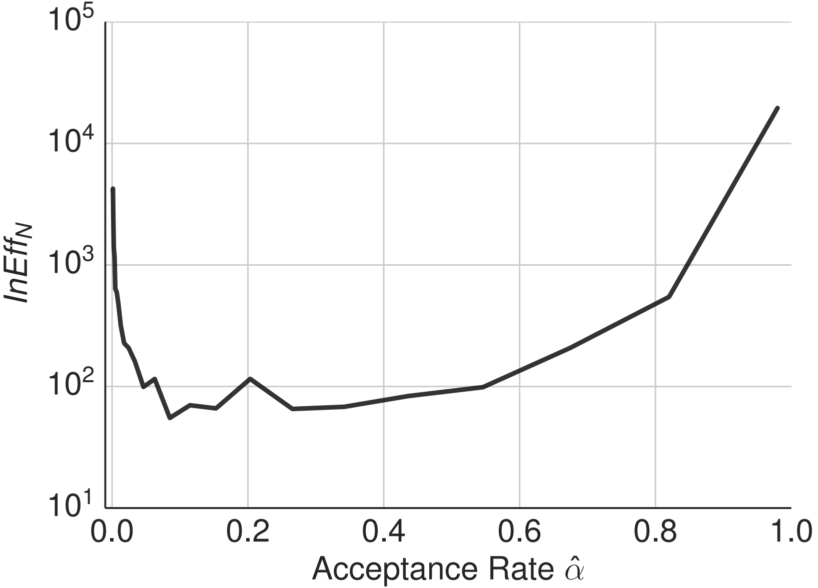
- Acceptance Rate
- $Ineff_\infty$
- $Ineff_N$

What is the relationship between acceptance rate and accuracy?

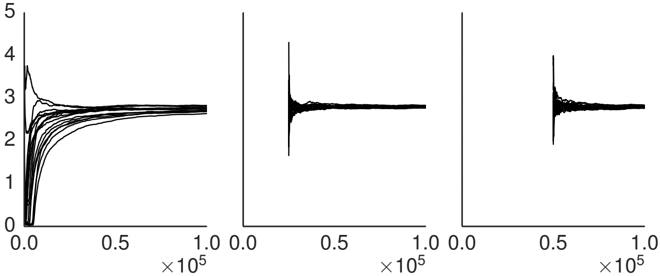
Effects of Scaling



Acceptance Rate vs. Accuracy



Convergence of Monte Carlo Average $\bar{\tau}_{N|N_0}$



Notes: The \$x-\$axis indicates the number of draws N. N_0 is set to 0, $25,000$ and $50,000$, respectively.

Improvements to MCMC: Blocking

- In high-dimensional parameter spaces the RWMH algorithm generates highly persistent Markov chains.
- What's bad about persistence?

$$\sqrt{N}(\bar{h}_N - \mathbb{E}[\bar{h}_N])$$

$$\Longrightarrow \quad N\left(0, \frac{1}{N}\sum_{i=1}^{n} \mathbb{V}[h(\theta^i)] + \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i} COV\big[h(\theta^i), h(\theta^j)\big]\right).$$

- Potential Remedy:
  - Partition $\theta = [\theta_1, \ldots, \theta_K]$.
  - Iterate over conditional posteriors $p(\theta_k | Y, \theta_{<-k>})$.
- To reduce persistence of the chain, try to find partitions such that parameters are strongly correlated within blocks and weakly correlated across blocks or use random blocking.

Block MH Algorithm

Draw $\theta^0 \in \Theta$ and then for $i = 1$ to N:

1. Create a partition B^i of the parameter

vector into N_{blocks} blocks $\theta_1, \ldots, \theta_{N_{blocks}}$ via some rule (perhaps probabilistic), unrelated to the current state of the Markov chain.

1. For $b = 1, \ldots, N_{blocks}$:

1. Draw $\vartheta_b \sim q(\cdot| \left[\theta^i_{<b}, \theta^{i-1}_b, \theta^{i-1}_{\geq b}\right])$.
2. With probability,

$$\alpha = \max\left\{\frac{p(\left[\theta^i_{<b}, \vartheta_b, \theta^{i-1}_{>b}\right]|Y)q(\theta^{i-1}_b, |\theta^i_{<b}, \vartheta_b, \theta^{i-1}_{>b})}{p(\theta^i_{<b}, \theta^{i-1}_b, \theta^{i-1}_{>b}|Y)q(\vartheta_b|\theta^i_{<b}, \theta^{i-1}_b, \theta^{i-1}_{>b})}, 1\right\},$$

set $\theta^i_b = \vartheta_b$, otherwise set $\theta^i_b = \theta^{i-1}_b$.

Random-Block MH Algorithm

- Generate a sequence of random partitions $\{B^i\}_{i=1}^N$ of the parameter vector θ into N_{blocks} equally sized blocks, denoted by θ_b, $b = 1, \ldots, N_{blocks}$ as follows:
  1. assign an $iidU[0, 1]$ draw to each element of θ;
  2. sort the parameters according to the assigned random number;
  3. let the b'th block consist of parameters $(b-1)N_{blocks}, \ldots, bN_{blocks}$.
- Execute Algorithm Block MH Algorithm.

Metropolis-Adjusted Langevin Algorithm

- The proposal distribution of Metropolis-Adjusted Langevin (MAL) algorithm is given by

$$\mu(\theta^{i-1}) = \theta^{i-1} + \frac{c_1}{2} M_1 \frac{\partial}{\partial\theta} \ln p(\theta^{i-1}|Y)\bigg|_{\theta=\theta^{i-1}},$$
$$\Sigma(\theta^{i-1}) = c_2^2 M_2.$$

  that is θ^{i-1} is adjusted by a step in the direction of the gradient of the log posterior density function.

- One standard practice is to set $M_1 = M_2 = M$, with

$$M = -\left[\frac{\partial}{\partial\theta\partial\theta'} \ln p(\theta|Y)\bigg|_{\theta=\hat{\theta}}\right]^{-1},$$

  where $\hat{\theta}$ is the mode of the posterior distribution obtained using a numerical optimization routine.

Newton MH Algorithm

- Newton MH Algorithm replaces the Hessian evaluated at the posterior mode $\hat{\theta}$ by the Hessian evaluated at θ^{i-1}.

- The proposal distribution is given by

$$\mu(\theta^{i-1}) = \theta^{i-1} - s \left[\frac{\partial}{\partial\theta\partial\theta'} \ln p(\theta|Y) \Big|_{\theta=\theta^{i-1}} \right]^{-1} \times \frac{\partial}{\partial\theta} \ln p(\theta^{i-1}|Y) \Big|_{\theta=\theta^{i-1}}$$

$$\hat{\Sigma}(\theta^{i-1}) = -c_2^2 \left[\frac{\partial}{\partial\theta\partial\theta'} \ln p(\theta|Y) \Big|_{\theta=\theta^{i-1}} \right]^{-1}.$$

- It is useful to let s be independently of θ^{i-1}:

$$c_1 = 2s, \quad s \sim iidU[0, \bar{s}],$$

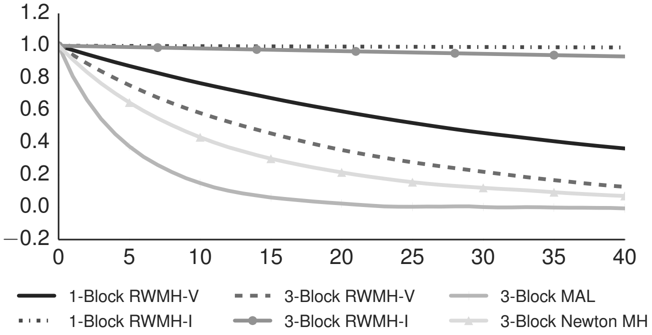
where $\bar{s}$ is a tuning parameter.

Run Times and Tuning Constants for MH Algorithms

Algorithm	Run Time [hh:mm:ss]	Accpt. Rate	Tuning Constants
1-Block RWMH-I	00:01:13	0.28	$c = 0.015$
1-Block RWMH-V	00:01:13	0.37	$c = 0.400$
3-Block RWMH-I	00:03:38	0.40	$c = 0.070$
3-Block RWMH-V	00:03:36	0.43	$c = 1.200$
3-Block MAL	00:54:12	0.43	$c_1 = 0.4, c_2 = 0.750$
3-Block Newton MH	03:01:40	0.53	$\bar{s} = 0.7, c_2 = 0.600$

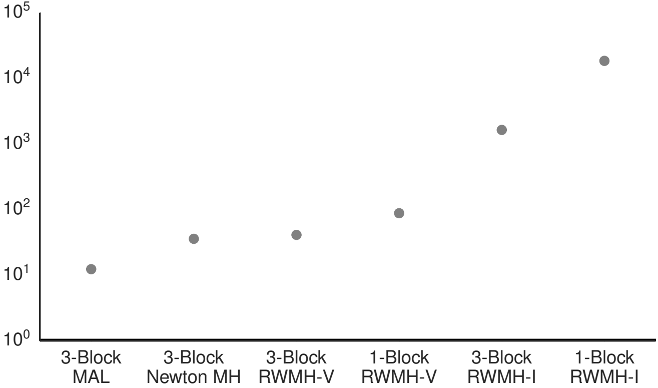
Notes: In each run we generate $N = 100,000$ draws. We report the fastest run time and the average acceptance rate across $N_{run} = 50$ independent Markov chains.

Autocorrelation Function of τ^i



Notes: The autocorrelation functions are computed based on a single run of each algorithm.

Inefficiency Factor $\text{InEff}_N[\bar{\tau}]$



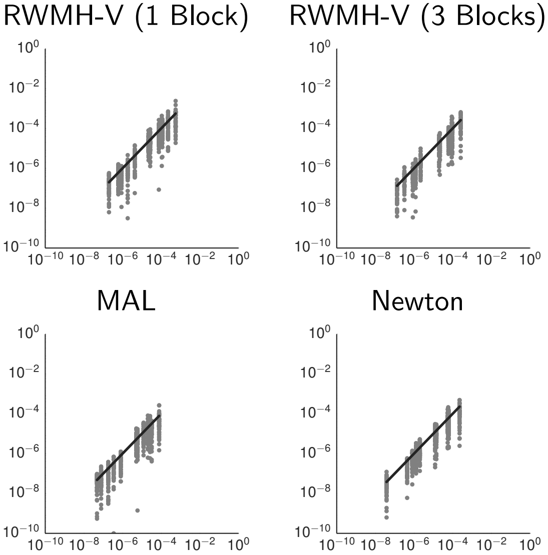
Notes: The small-sample inefficiency factors are computed based on $N_{run} = 50$ independent runs of each algorithm.

IID Equivalent Draws Per Second

$$\textit{iid}\text{-equivalent draws per second} = \frac{N}{\text{Run Time [seconds]}} \cdot \frac{1}{\text{InEff}_N}.$$

Algorithm	Draws Per Second
1-Block RWMH-V	7.76
3-Block RWMH-V	5.65
3-Block MAL	1.24
3-Block RWMH-I	0.14
3-Block Newton MH	0.13
1-Block RWMH-I	0.04

Performance of Different MH Algorithms



Notes: Each panel contains scatter plots of the small sample variance $\mathbb{V}[\bar{\theta}]$ computed across multiple chains (x-axis) versus the HAC$[\bar{h}]$ estimates of $\Omega(\theta)/N$ (y-axis).

Recall: Posterior Odds and Marginal Data Densities

- Posterior model probabilities can be computed as follows:

$$\pi_{i,T} = \frac{\pi_{i,0} p(Y|\mathcal{M}_i)}{\sum_j \pi_{j,0} p(Y|\mathcal{M}_j)}, \quad j = 1, \dots, 2, \tag{6}$$

where

$$p(Y|\mathcal{M}) = \int p(Y|\theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta \tag{7}$$

- Note:

$$\ln p(Y_{1:T}|\mathcal{M}) = \sum_{t=1}^{T} \ln \int p(y_t|\theta, Y_{1:t-1}, \mathcal{M}) p(\theta|Y_{1:t-1}, \mathcal{M}) d\theta$$

- Posterior odds and Bayes Factor

$$\frac{\pi_{1,T}}{\pi_{2,T}} = \underbrace{\frac{\pi_{1,0}}{\pi_{2,0}}}_{Prior\ Odds} \times \underbrace{\frac{p(Y|\mathcal{M}_1)}{p(Y|\mathcal{M}_2)}}_{Bayes\ Factor} \tag{8}$$

Computation of Marginal Data Densities

- Reciprocal importance sampling:
  - Geweke's modified harmonic mean estimator
  - Sims, Waggoner, and Zha's estimator
- Chib and Jeliazkov's estimator
- For a survey, see Ardia, Hoogerheide, and van Dijk (2009).

Modified Harmonic Mean

- Reciprocal importance samplers are based on the following identity:

$$\frac{1}{p(Y)} = \int \frac{f(\theta)}{p(Y|\theta)p(\theta)} p(\theta|Y)d\theta, \tag{9}$$

  where $\int f(\theta)d\theta = 1$.

- Conditional on the choice of $f(\theta)$ an obvious estimator is

$$\hat{p}_G(Y) = \left[\frac{1}{N}\sum_{i=1}^{N} \frac{f(\theta^i)}{p(Y|\theta^i)p(\theta^i)}\right]^{-1}, \tag{10}$$

  where θ^i is drawn from the posterior $p(\theta|Y)$.

- Geweke (1999):

$$\begin{aligned} f(\theta) &= \tau^{-1}(2\pi)^{-d/2}|V_\theta|^{-1/2}\exp\left[-0.5(\theta-\bar{\theta})'V_\theta^{-1}(\theta-\bar{\theta})\right] \\ &\quad \times \left\{(\theta-\bar{\theta})'V_\theta^{-1}(\theta-\bar{\theta}) \leq F^{-1}_{\chi^2_d}(\tau)\right\}. \end{aligned} \tag{11}$$

Chib and Jeliazkov

- Rewrite Bayes Theorem:

$$p(Y) = \frac{p(Y|\theta)p(\theta)}{p(\theta|Y)}. \tag{12}$$

- Thus,

$$\hat{p}_{CS}(Y) = \frac{p(Y|\tilde{\theta})p(\tilde{\theta})}{\hat{p}(\tilde{\theta}|Y)}, \tag{13}$$

  where we replaced the generic θ in~(12) by the posterior mode $\tilde{\theta}$.

Chib and Jeliazkov

- Use output of Metropolis-Hastings Algorithm.
- Proposal density for transition $\theta \mapsto \tilde{\theta}$: $q(\theta, \tilde{\theta}|Y)$.
- Probability of accepting proposed draw:

$$\alpha(\theta, \tilde{\theta}|Y) = \min \left\{ 1, \ \frac{p(\tilde{\theta}|Y)/q(\theta, \tilde{\theta}|Y)}{p(\theta|Y)/q(\tilde{\theta}, \theta|Y)} \right\}.$$

- Note that

$$\begin{aligned}
&\int \alpha(\theta, \tilde{\theta}|Y) q(\theta, \tilde{\theta}|Y) p(\theta|Y) d\theta \\
&= \int \min \left\{ 1, \ \frac{p(\tilde{\theta}|Y)/q(\theta, \tilde{\theta}|Y)}{p(\theta|Y)/q(\tilde{\theta}, \theta|Y)} \right\} q(\theta, \tilde{\theta}|Y) p(\theta|Y) d\theta \\
&= p(\tilde{\theta}|Y) \int \min \left\{ \frac{p(\theta|Y)/q(\tilde{\theta}, \theta|Y)}{p(\tilde{\theta}|Y)/q(\theta, \tilde{\theta}|Y)}, \ 1 \right\} q(\tilde{\theta}, \theta|Y) d\theta \\
&= p(\tilde{\theta}|Y) \int \alpha(\tilde{\theta}, \theta|Y) q(\tilde{\theta}, \theta|Y) d\theta
\end{aligned}$$

Chib and Jeliazkov

- Posterior density at the mode can be approximated as follows

$$\hat{p}(\tilde{\theta}|Y) = \frac{\frac{1}{N}\sum_{i=1}^{N}\alpha(\theta^{i},\tilde{\theta}|Y)q(\theta^{i},\tilde{\theta}|Y)}{\frac{1}{J}\sum_{j=1}^{J}\alpha(\tilde{\theta},\theta^{j}|Y)}, \tag{14}$$

- $\{\theta^i\}$ are posterior draws obtained with the the M-H Algorithm;
- $\{\theta^j\}$ are additional draws from $q(\tilde{\theta},\theta|Y)$ given the fixed value $\tilde{\theta}$.

MH-Based Marginal Data Density Estimates

Model	Mean($\ln \hat{p}(Y)$)	Std. Dev.($\ln \hat{p}(Y)$)
Geweke ($\tau = 0.5$)	-346.17	0.03
Geweke ($\tau = 0.9$)	-346.10	0.04
SWZ ($q = 0.5$)	-346.29	0.03
SWZ ($q = 0.9$)	-346.31	0.02
Chib and Jeliazkov	-346.20	0.40

Notes: Table shows mean and standard deviation of log marginal data density estimators, computed over $N_{run} = 50$ runs of the RWMH-V sampler using $N = 100,000$ draws, discarding a burn-in sample of $N_0 = 50,000$ draws. The SWZ estimator uses $J = 100,000$ draws to compute $\hat{\tau}$, while the CJ estimators uses $J = 100,000$ to compute the denominator of $\hat{p}(\tilde{\theta}|Y)$.

References

CHIB, S. AND E. GREENBERG (1995): "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

DEL NEGRO, M. AND F. SCHORFHEIDE (2008): "Forming Priors for DSGE Models (and How it Affects the Assessment of Nominal Rigidities)," *Journal of Monetary Economics*, 55, 1191–1208.

GEWEKE, J. (2005): *Contemporary Bayesian Econometrics and Statistics*, John Wiley & Sons, Inc.

HASTINGS, W. (1970): "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, AND E. TELLER (1953): "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.

ROBERT, C. P. AND G. CASELLA (2004): *Monte Carlo Statistical Methods*, Springer.

ROBERTS, G., A. GELMAN, AND G. W.R. (1997): "Weak