Bayesian Estimation of DSGE Models

Frank Schorfheide
University of Pennsylvania
Gerzensee Ph.D. Course on Bayesian Macroeconometrics

May 28, 2019
Bayesian Inference

- **Ingredients of Bayesian Analysis:**
  - Likelihood function \( p(Y|\theta) \)
  - Prior density \( p(\theta) \)
  - Marginal data density \( p(Y) = \int p(Y|\theta)p(\theta)d\phi \)

- **Bayes Theorem:**
  \[
p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \propto p(Y|\theta)p(\theta)
  \]

- **Implementation:** usually by generating a sequence of draws (not necessarily iid) from posterior
  \[
  \theta^i \sim p(\theta|Y), \quad i = 1, \ldots, N
  \]

- **Algorithms:** direct sampling, accept/reject sampling, importance sampling, Markov chain Monte Carlo sampling, sequential Monte Carlo sampling...
Generic Metropolis-Hastings Algorithm

For \( i = 1 \) to \( N \):

1. Draw \( \vartheta \) from a density \( q(\vartheta|\theta^{i-1}) \).
2. 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.

Recall \( p(\theta|Y) \propto p(Y|\theta)p(\theta) \).

We draw \( \theta^i \) conditional on a parameter draw \( \theta^{i-1} \): leads to Markov transition kernel \( K(\theta|\tilde{\theta}) \).
Why Does it Work?

- Algorithm generates a Markov transition kernel $K(\theta | \tilde{\theta})$: it takes a draw $\theta^{i-1}$ and uses some randomization to turn it into a draw $\theta^i$.

- Important invariance property: if $\theta^{i-1}$ is from posterior $p(\theta | Y)$, then $\theta^i$’s distribution will also be $p(\theta | Y)$.

- Contraction property: if $\theta^{i-1}$ is from some distribution $\pi_{i-1}(\theta)$, then the discrepancy between the “true” posterior and

$$
\pi_i(\theta) = \int K(\theta | \tilde{\theta})\pi_{i-1}(\tilde{\theta})d\tilde{\theta}
$$

is smaller than the discrepancy between $\pi_{i-1}(\theta)$ and $p(\theta | Y)$. 
The Invariance Property

- It can be shown that
  \[ p(\theta | Y) = \int K(\theta | \tilde{\theta}) p(\tilde{\theta} | Y) d\tilde{\theta}. \]

- Write
  \[ K(\theta | \tilde{\theta}) = u(\theta | \tilde{\theta}) + r(\tilde{\theta}) \delta_{\tilde{\theta}}(\theta). \]

- \( u(\theta | \tilde{\theta}) \) is the density kernel (note that \( u(\theta | \cdot) \) does not integrate to one) for accepted draws:
  \[ u(\theta | \tilde{\theta}) = \alpha(\theta | \tilde{\theta}) q(\theta | \tilde{\theta}). \]

- Rejection probability:
  \[ r(\tilde{\theta}) = \int [1 - \alpha(\theta | \tilde{\theta})] q(\theta | \tilde{\theta}) d\theta = 1 - \int u(\theta | \tilde{\theta}) d\theta. \]
The Invariance Property

• Reversibility: Conditional on the sampler not rejecting the proposed draw, the density associated with a transition from $\tilde{\theta}$ to $\theta$ is identical to the density associated with a transition from $\theta$ to $\tilde{\theta}$:

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

$$= \min \left\{ p(\tilde{\theta} | Y) q(\theta | \tilde{\theta}), p(\theta | Y) q(\tilde{\theta} | \theta) \right\}$$

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

$$= p(\theta | Y) u(\tilde{\theta} | \theta).$$

• Using the reversibility result, we can now verify the invariance property:

$$\int K(\theta | \tilde{\theta}) p(\tilde{\theta} | Y) d\tilde{\theta} = \int u(\theta | \tilde{\theta}) p(\tilde{\theta} | Y) d\tilde{\theta} + \int r(\tilde{\theta}) \delta_{\tilde{\theta}}(\theta) p(\tilde{\theta} | Y) d\tilde{\theta}$$

$$= \int u(\tilde{\theta} | \theta) p(\theta | Y) d\tilde{\theta} + r(\theta) p(\theta | Y)$$

$$= p(\theta | Y)$$
• Suppose parameter vector $\theta$ is scalar and takes only two values:

$$\Theta = \{\tau_1, \tau_2\}$$

• The posterior distribution $p(\theta|Y)$ can be represented by a set of probabilities collected in the vector $\pi$, say $\pi = [\pi_1, \pi_2]$ with $\pi_2 > \pi_1$.

• Suppose we obtain $\vartheta$ based on transition matrix $Q$:

$$Q = \begin{bmatrix} q & (1-q) \\ (1-q) & q \end{bmatrix}.$$
Example: Discrete MH Algorithm

• Iteration $i$: suppose that $\theta^{i-1} = \tau_j$. Based on transition matrix

$$Q = \begin{bmatrix} q & (1-q) \\ (1-q) & q \end{bmatrix},$$

determine a proposed state $\vartheta = \tau_s$.

• With probability $\alpha(\tau_s | \tau_j)$ the proposed state is accepted. Set $\theta^i = \vartheta = \tau_s$.

• With probability $1 - \alpha(\tau_s | \tau_j)$ stay in old state and set $\theta^i = \theta^{i-1} = \tau_j$.

• Choose ($Q$ terms cancel because of symmetry)

$$\alpha(\tau_s | \tau_j) = \min \left\{ 1, \frac{\pi_s}{\pi_j} \right\}.$$
Example: Transition Matrix

- The resulting chain’s transition matrix is:

\[ K = \begin{bmatrix} q & (1 - q) \\ (1 - q)\frac{\pi_1}{\pi_2} & q + (1 - q)\left(1 - \frac{\pi_1}{\pi_2}\right) \end{bmatrix}. \]

- Straightforward calculations reveal that the transition matrix \( K \) has eigenvalues:

\[ \lambda_1(K) = 1, \quad \lambda_2(K) = q - (1 - q)\frac{\pi_1}{1 - \pi_1}. \]

- Equilibrium distribution is eigenvector associated with unit eigenvalue.

- For \( q \in [0, 1) \) the equilibrium distribution is unique.
Example: Convergence

- The persistence of the Markov chain depends on second eigenvalue, which depends on the proposal distribution $Q$.

- Define the transformed parameter

$$\xi^i = \frac{\theta^i - \tau_1}{\tau_2 - \tau_1}.$$ 

- We can represent the Markov chain associated with $\xi^i$ as first-order autoregressive process

$$\xi^i = (1 - k_{22}) + \lambda_2(K)\xi^{i-1} + \nu^i.$$ 

- Conditional on $\xi^i = j$, $j = 0, 1$, the innovation $\nu^i$ has support on $k_{jj}$ and $(1 - k_{jj})$, its conditional mean is equal to zero, and its conditional variance is equal to $k_{jj}(1 - k_{jj})$. 

Example: Convergence

- Autocovariance function of $h(\theta^i)$:
  \[
  COV(h(\theta^i), h(\theta^{(i-l)})) = \frac{(h(\tau_2) - h(\tau_1))^2 \pi_1 (1 - \pi_1)}{\pi_1 (1 - \pi_1)} \left( q - (1 - q) \frac{\pi_1}{1 - \pi_1} \right)^l
  \]

- If $q = \pi_1$ then the autocovariances are equal to zero and the draws $h(\theta^i)$ are serially uncorrelated (in fact, in our simple discrete setting they are also independent).
Example: Convergence

- Define the Monte Carlo estimate

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

- Deduce from CLT

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

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

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

- In turn, the asymptotic inefficiency factor is given by

\[ \text{InEff}_\infty = \frac{\Omega(h)}{\mathbb{V}_\pi[h]} = 1 + 2 \lim_{L \to \infty} \sum_{l=1}^{L} \frac{L - l}{L} \left( q - (1 - q) \frac{\pi_1}{1 - \pi_1} \right)^l. \]
Example: Autocorrelation Function of $\theta^i$, $\pi_1 = 0.2$
Example: Asymptotic Inefficiency $\text{InEff}_\infty$, $\pi_1 = 0.2$
Example: Small Sample Variance $\mathbb{V}[\hat{h}_N]$ versus HAC Estimates of $\Omega(h)$
Benchmark Random-Walk Metropolis-Hastings (RWMH) Algorithm for DSGE Models

• Initialization:
  1. Use a numerical optimization routine to maximize the log posterior, which up to a constant is given by \( \ln p(Y|\theta) + \ln p(\theta) \). Denote the posterior mode by \( \hat{\theta} \).
  2. Let \( \hat{\Sigma} \) be the inverse of the (negative) Hessian computed at the posterior mode \( \hat{\theta} \), which can be computed numerically.
  3. Draw \( \theta^0 \) from \( N(\hat{\theta}, c_0^2 \hat{\Sigma}) \) or directly specify a starting value.

• Main Algorithm – For \( i = 1, \ldots, N \):
  1. Draw \( \vartheta \) from the proposal distribution \( N(\theta^{i-1}, c^2 \hat{\Sigma}) \).
  2. Set \( \theta^i = \vartheta \) with probability

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

and \( \theta^i = \theta^{i-1} \) otherwise.
Notes: Output growth per capita is measured in quarter-on-quarter (Q-o-Q) percentages. Inflation is CPI inflation in annualized Q-o-Q percentages. Federal funds rate is the average annualized effective funds rate for each quarter.
Convergence of Monte Carlo Average $\bar{\tau}_{N|N_0}$
### Posterior Estimates of DSGE Model Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>[0.05, 0.95]</th>
<th>Parameter</th>
<th>Mean</th>
<th>[0.05,0.95]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>2.83</td>
<td>[ 1.95, 3.82]</td>
<td>$\rho_r$</td>
<td>0.77</td>
<td>[ 0.71, 0.82]</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.78</td>
<td>[ 0.51, 0.98]</td>
<td>$\rho_g$</td>
<td>0.98</td>
<td>[ 0.96, 1.00]</td>
</tr>
<tr>
<td>$\psi_1$</td>
<td>1.80</td>
<td>[ 1.43, 2.20]</td>
<td>$\rho_z$</td>
<td>0.88</td>
<td>[ 0.84, 0.92]</td>
</tr>
<tr>
<td>$\psi_2$</td>
<td>0.63</td>
<td>[ 0.23, 1.21]</td>
<td>$\sigma_r$</td>
<td>0.22</td>
<td>[ 0.18, 0.26]</td>
</tr>
<tr>
<td>$r^{(A)}$</td>
<td>0.42</td>
<td>[ 0.04, 0.95]</td>
<td>$\sigma_g$</td>
<td>0.71</td>
<td>[ 0.61, 0.84]</td>
</tr>
<tr>
<td>$\pi^{(A)}$</td>
<td>3.30</td>
<td>[ 2.78, 3.80]</td>
<td>$\sigma_z$</td>
<td>0.31</td>
<td>[ 0.26, 0.36]</td>
</tr>
<tr>
<td>$\gamma^{(Q)}$</td>
<td>0.52</td>
<td>[ 0.28, 0.74]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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.
Notes: Results are based on \( N_{\text{run}} = 50 \) independent Markov chains. The acceptance rate (average across multiple chains), HAC-based estimate of \( \text{InEff}_{\infty}[\bar{\tau}] \) (average across multiple chains), and \( \text{InEff}_{N}[\bar{\tau}] \) are shown as a function of the scaling constant \( c \).
Notes: $\text{InEff}_N[\tilde{\tau}]$ versus the acceptance rate $\hat{\alpha}$.
What Can We Do With Our Posterior Draws?

- Store them on our harddrive!
- Convert them into objects of interest:
  - impulse response functions;
  - government spending multipliers;
  - welfare effects of target inflation rate changes;
  - forecasts;
  - (...)
Parameter Transformations: Impulse Responses

Notes: The figure depicts pointwise posterior means and 90% credible bands. The responses of output are in percent relative to the initial level, whereas the responses of inflation and interest rates are in annualized percentages.
• The posterior expected loss of decision $\delta(\cdot)$:

$$\rho(\delta(\cdot)|Y) = \int_\Theta L(\theta, \delta(Y)) p(\theta|Y) d\theta.$$  

• Bayes decision minimizes the posterior expected loss:

$$\delta^*(Y) = \arg\min_d \rho(\delta(\cdot)|Y).$$  

• Approximate $\rho(\delta(\cdot)|Y)$ by a Monte Carlo average

$$\bar{\rho}_N(\delta(\cdot)|Y) = \frac{1}{N} \sum_{i=1}^{N} L(\theta^i, \delta(\cdot)).$$  

• Then compute

$$\delta^*_N(Y) = \arg\min_d \bar{\rho}_N(\delta(\cdot)|Y).$$
• Consider the following identity:

\[
\frac{1}{p(Y)} = \int \frac{f(\theta)}{p(Y|\theta)p(\theta)} p(\theta|Y) d\theta,
\]

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

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

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

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

• Geweke (1999):

\[
f(\theta) = \tau^{-1}(2\pi)^{-d/2}|V_\theta|^{-1/2} \exp \left[ -0.5(\theta - \bar{\theta})' V_\theta^{-1}(\theta - \bar{\theta}) \right] \\
\times \left\{ (\theta - \bar{\theta})' V_\theta^{-1}(\theta - \bar{\theta}) \leq F_{\chi^2_\tau}^{-1}(\tau) \right\}.
\]
Challenges Due to Irregular Posteriors

A stylized state-space model:

\[ y_t = [1 \ 1]s_t, \quad s_t = \begin{bmatrix} \phi_1 & 0 \\ \phi_3 & \phi_2 \end{bmatrix} s_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon_t, \quad \epsilon_t \sim iidN(0, 1). \]

where

- Structural parameters \( \theta = [\theta_1, \theta_2]' \), domain is unit square.
- Reduced-form parameters \( \phi = [\phi_1, \phi_2, \phi_3]' \)

\[ \phi_1 = \theta_1^2, \quad \phi_2 = (1 - \theta_1^2), \quad \phi_3 - \phi_2 = -\theta_1 \theta_2. \]
Challenges Due to Irregular Posteriors

- $s_{1,t}$ looks like an exogenous technology process.
- $s_{2,t}$ evolves like an endogenous state variable, e.g., the capital stock.
- $\theta_2$ is not identifiable if $\theta_1 = 0$ because $\theta_2$ enters the model only multiplicatively.
- Law of motion of $y_t$ is restricted ARMA$(2,1)$ process:

\[
(1 - \theta_1^2 L)(1 - (1 - \theta_1^2)L)y_t = (1 - \theta_1 \theta_2 L)\epsilon_t.
\]

- Given $\theta_1$ and $\theta_2$, we obtain an observationally equivalent process by switching the values of the two roots of the autoregressive lag polynomial.
- Choose $\tilde{\theta}_1$ and $\tilde{\theta}_2$ such that

\[
\tilde{\theta}_1 = \sqrt{1 - \theta_1^2}, \quad \tilde{\theta}_2 = \theta_1 \theta_2 / \tilde{\theta}_1.
\]
Notes: Intersections of the solid lines indicate parameter values that were used to generate the data from which the posteriors are constructed. Left panel: $\theta_1 = 0.1$ and $\theta_2 = 0.5$. Right panel: $\theta_1 = 0.8$, $\theta_2 = 0.3$. 
• In high-dimensional parameter spaces the RWMH algorithm generates highly persistent Markov chains.

• What’s bad about persistence?

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

$$\implies N\left(0, \frac{1}{N} \sum_{i=1}^{n} \nabla[h(\theta^i)] + \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \text{COV}[h(\theta^i), h(\theta^j)]\right).$$

• 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.
Improvements to MCMC: Blocking

- Chib and Ramamurthy (2010, JoE):
  - Use randomized partitions
  - Use simulated annealing to find mode of $p(\theta_k | Y, \theta_{<k})$. Then construct Hessian to obtain
    covariance matrix for proposal density.

- Herbst (2011, Penn Dissertation):
  - Utilize analytical derivatives
  - Use information in Hessian (evaluated at an earlier parameter draw) to construct parameter
    blocks. For non-elliptical distribution partitions change as sampler moves through parameter
    space.
  - Use Gauss-Newton step to construct proposal densities
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 $N_{\text{blocks}}$ blocks $\theta_1, \ldots, \theta_{N_{\text{blocks}}}$ via some rule (perhaps probabilistic), unrelated to the current state of the Markov chain.

2. For $b = 1, \ldots, N_{\text{blocks}}$:
   1. Draw $\vartheta_b \sim q(\cdot [\theta^i_{<b}, \theta^i_{b-1}, \theta^i_{\geq b}])$.
   2. With probability,

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

   set $\theta^i_b = \vartheta_b$, otherwise set $\theta^i_b = \theta^i_{b-1}$. 

Frank Schorfheide  
Bayesian Estimation of DSGE Models
Random-Block MH Algorithm

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

2 Execute Algorithm Block MH Algorithm.

---

\(^1\)If the number of parameters is not divisible by \( N_{\text{blocks}} \), then the size of a subset of the blocks has to be adjusted.
Run Times and Tuning Constants for MH Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Run Time [hh:mm:ss]</th>
<th>Acceptance Rate</th>
<th>Tuning Constants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Block RWMH-I</td>
<td>00:01:13</td>
<td>0.28</td>
<td>$c = 0.015$</td>
</tr>
<tr>
<td>1-Block RWMH-V</td>
<td>00:01:13</td>
<td>0.37</td>
<td>$c = 0.400$</td>
</tr>
<tr>
<td>3-Block RWMH-I</td>
<td>00:03:38</td>
<td>0.40</td>
<td>$c = 0.070$</td>
</tr>
<tr>
<td>3-Block RWMH-V</td>
<td>00:03:36</td>
<td>0.43</td>
<td>$c = 1.200$</td>
</tr>
<tr>
<td>3-Block MAL</td>
<td>00:54:12</td>
<td>0.43</td>
<td>$c_1 = 0.400, c_2 = 0.750$</td>
</tr>
<tr>
<td>3-Block Newton MH</td>
<td>03:01:40</td>
<td>0.53</td>
<td>$\bar{c} = 0.700, c_2 = 0.600$</td>
</tr>
</tbody>
</table>

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. See book for MAL and Newton MH Algorithms.
Notes: The autocorrelation functions are computed based on a single run of each algorithm.
Notes: The small sample inefficiency factors are computed based on $N_{\text{run}} = 50$ independent runs of each algorithm.
IID Equivalent Draws Per Second

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

- 3-Block MAL: 1.24
- 3-Block Newton MH: 0.13
- 3-Block RWMH-V: 5.65
- 1-Block RWMH-V: 7.76
- 3-Block RWMH-I: 0.14
- 1-Block RWMH-I: 0.04
Performance of Different MH Algorithms

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