Bayesian Inference

Frank Schorfheide
University of Pennsylvania
Econ 722 – Part 1

February 13, 2019

Statistical Inference

- Frequentist:
    - pre-experimental perspective;
    - condition on "true" but unknown θ_0;
    - treat data Y as random;
    - study behavior of estimators and decision rules under repeated sampling.
- Bayesian:
    - post-experimental perspective;
    - condition on observed sample Y;
    - treat parameter θ as unknown and random;
    - derive estimators and decision rules that minimize expected loss (averaging over θ) conditional on observed Y.

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

$$\theta^i \sim p(\theta|Y), \quad i = 1, \dots, N$$

- Algorithms: direct sampling, accept/reject sampling, importance sampling, Markov chain Monte Carlo sampling, sequential Monte Carlo sampling...

- We previously discussed the evaluation of the likelihood function: given a parameter θ
  - solve the DSGE model to obtain the state-space representation;
  - use the Kalman filter to evaluate the likelihood function.
- Let's talk a bit about prior distributions.

Prior Distributions

- Ideally: probabilistic representation of our knowledge/beliefs before observing sample Y.
- More realistically: choice of prior as well as model are influenced by some observations. Try to keep influence small or adjust measures of uncertainty.
- Views about role of priors:
  1. keep them "uninformative" (???) so that posterior inherits shape of likelihood function;
  2. use them to regularize the likelihood function;
  3. incorporate information from sources other than Y;

Prior Elicitation for DSGE Models

- Group parameters:
  - steady-state related parameters
  - parameters assoc with exogenous shocks
  - parameters assoc with internal propagation
- Non-sample information $p(\theta|\mathcal{X}^0)$:
  - pre-sample information
  - micro-level information
- To guide the prior for θ, you can ask: what are its implications for observables Y?

Prior Distribution

Name	Domain	Prior		
		Density	Para (1)	Para (2)
Steady-State-Related Parameters $\theta_{(ss)}$				
$100(1/\beta - 1)$	$\mathbb{R}^+$	Gamma	0.50	0.50
$100 \log \pi^*$	$\mathbb{R}^+$	Gamma	1.00	0.50
$100 \log \gamma$	$\mathbb{R}$	Normal	0.75	0.50
λ	$\mathbb{R}^+$	Gamma	0.20	0.20
Endogenous Propagation Parameters $\theta_{(endo)}$				
ζ_p	$[0, 1]$	Beta	0.70	0.15
$1/(1+\nu)$	$\mathbb{R}^+$	Gamma	1.50	0.75

Notes: Marginal prior distributions for each DSGE model parameter. Para (1) and Para (2) list the means and the standard deviations for Beta, Gamma, and Normal distributions; the upper and lower bound of the support for the Uniform distribution; s and ν for the Inverse Gamma distribution, where $p_{\mathcal{IG}}(\sigma|\nu, s) \propto \sigma^{-\nu-1} e^{-\nu s^2/2\sigma^2}$. The joint prior distribution of θ is truncated at the boundary of the determinacy region.

Prior Distribution

Name	Domain	Prior		
		Density	Para (1)	Para (2)
		Exogenous Shock Parameters $\theta_{(exo)}$		
ρ_ϕ	$[0,1)$	Uniform	0.00	1.00
ρ_λ	$[0,1)$	Uniform	0.00	1.00
ρ_z	$[0,1)$	Uniform	0.00	1.00
$100\sigma_\phi$	$\mathbb{R}^+$	InvGamma	2.00	4.00
$100\sigma_\lambda$	$\mathbb{R}^+$	InvGamma	0.50	4.00
$100\sigma_z$	$\mathbb{R}^+$	InvGamma	2.00	4.00
$100\sigma_r$	$\mathbb{R}^+$	InvGamma	0.50	4.00

Notes: Marginal prior distributions for each DSGE model parameter. Para (1) and Para (2) list the means and the standard deviations for Beta, Gamma, and Normal distributions; the upper and lower bound of the support for the Uniform distribution; s and ν for the Inverse Gamma distribution, where $p_{\mathcal{IG}}(\sigma|\nu,s) \propto \sigma^{-\nu-1}e^{-\nu s^2/2\sigma^2}$. The joint prior distribution of θ is truncated at the boundary of the determinacy region.

Draws from Posterior

- We will focus on Markov chain Monte Carlo (MCMC) algorithms that generate draws $\{\theta^i\}_{i=1}^N$ from posterior distributions of parameters.
- Draws can then be transformed into objects of interest, $h(\theta^i)$, and under suitable conditions a Monte Carlo average of the form

$$\bar{h}_N = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i) \approx \mathbb{E}_\pi[h] = \int h(\theta)p(\theta|Y)d\theta.$$

- Strong law of large numbers (SLLN), central limit theorem (CLT)...

Markov Chain Monte Carlo (MCMC)

- Main idea: create a sequence of serially correlated draws such that the distribution of θ^i converges to the posterior distribution $p(\theta|Y)$.

Generic Metropolis-Hastings Algorithm

For $i = 1$ to N:

1. Draw ϑ 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 θ^i conditional on a parameter draw θ^{i-1}: leads to Markov transition kernel $K(\theta|\tilde{\theta})$.

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 θ^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 ϑ 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.

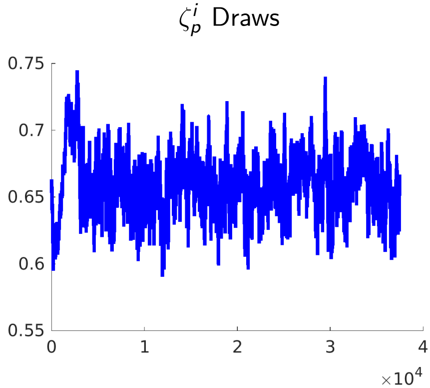
Benchmark RWMH Algorithm for DSGE Models

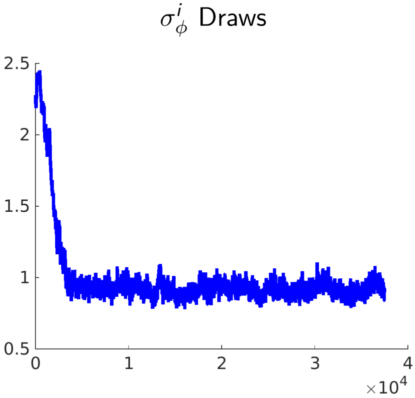
- Initialization steps can be modified as needed for particular application.
- If numerical optimization does not work well, one could let $\hat{\Sigma}$ be a diagonal matrix with prior variances on the diagonal.
- Or, $\hat{\Sigma}$ could be based on a preliminary run of a posterior sampler.
- It is good practice to run multiple chains based on different starting values.

Numerical Illustration

- Generate a single sample of size $T = 80$ from the stylized DSGE model.
- Combine likelihood and prior to form posterior.
- Draws from this posterior distribution are generated using the RWMH algorithm.
- Chain is initialized with a draw from the prior distribution.
- The covariance matrix $\hat{\Sigma}$ is based on the negative inverse Hessian at the mode. The scaling constant c is set equal to 0.075, which leads to an acceptance rate for proposed draws of 0.55.

Parameter Draws from MH Algorithm

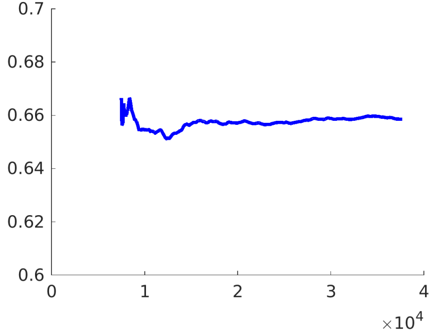




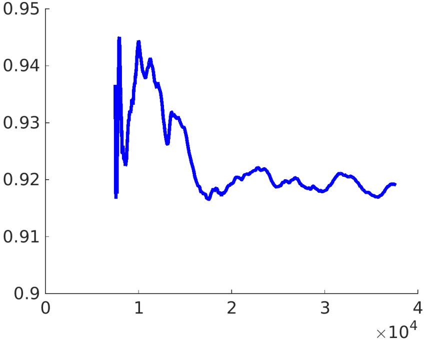
Notes: The posterior is based on a simulated sample of observations of size $T = 80$. The top panel shows the sequence of parameter draws and the bottom panel shows recursive means.

Parameter Draws from MH Algorithm

Recursive Mean $\frac{1}{N-N_0}\sum_{i=N_0+1}^{N}\zeta_p^i$

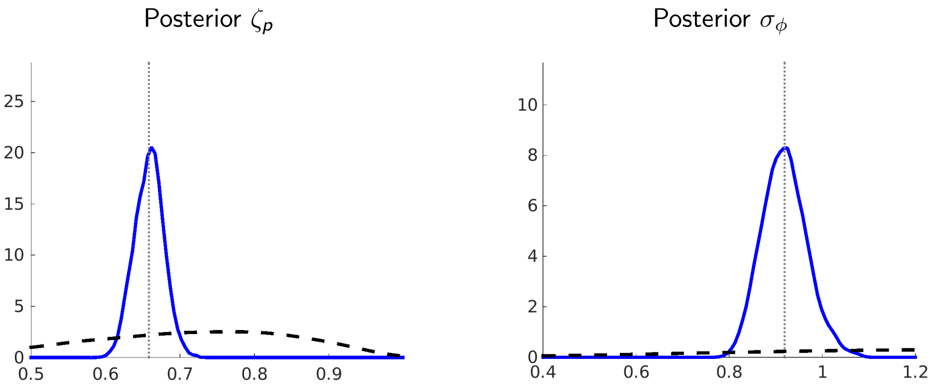


Recursive Mean $\frac{1}{N-N_0}\sum_{i=N_0+1}^{N}\sigma_\phi^i$



Notes: The posterior is based on a simulated sample of observations of size $T = 80$. The top panel shows the sequence of parameter draws and the bottom panel shows recursive means.

Prior and Posterior Densities



Notes: The dashed lines represent the prior densities, whereas the solid lines correspond to the posterior densities of ζ_p and σ_ϕ. The posterior is based on a simulated sample of observations of size $T = 80$. We generate $N = 37,500$ draws from the posterior and drop the first $N_0 = 7,500$ draws.

Why Does it Work?

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

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

- Contraction property: if θ^{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 integrated to one) for accepted draws:

$$u(\theta|\tilde{\theta}) = \alpha(\theta|\tilde{\theta}) q(\theta|\tilde{\theta}).$$

- Rejection probability:

$$r(\tilde{\theta}) = \int \left[1 - \alpha(\theta|\tilde{\theta})\right] q(\theta|\tilde{\theta}) d\theta = 1 - \int u(\theta|\tilde{\theta}) d\theta.$$

The Invariance Property

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

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

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

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

A Discrete Example

- Suppose parameter vector θ 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 π, say $\pi = [\pi_1, \pi_2]$ with $\pi_2 > \pi_1$.
- Suppose we obtain ϑ 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:

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

- 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 ξ^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 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 $k_{jj}(1 - k_{jj})$.

Example: Convergence

- Autocovariance function of $h(\theta^i)$:

$$
\begin{aligned}
&COV\big(h(\theta^i), h(\theta^{(i-l)})\big) \\
&\quad = \quad \big(h(\tau_2) - h(\tau_1)\big)^2 \pi_1(1-\pi_1)\left(q - (1-q)\frac{\pi_1}{1-\pi_1}\right)^l \\
&\quad = \quad \mathbb{V}_\pi[h]\left(q - (1-q)\frac{\pi_1}{1-\pi_1}\right)^l
\end{aligned}
$$

- 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

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

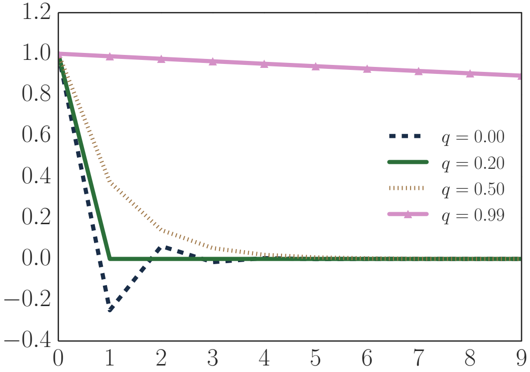
where $\Omega(h)$ is the long-run covariance matrix

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

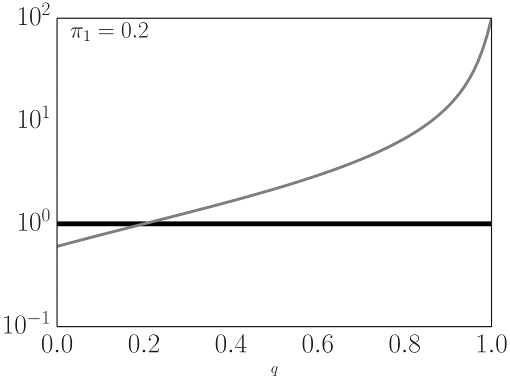
- In turn, the asymptotic inefficiency factor is given by

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

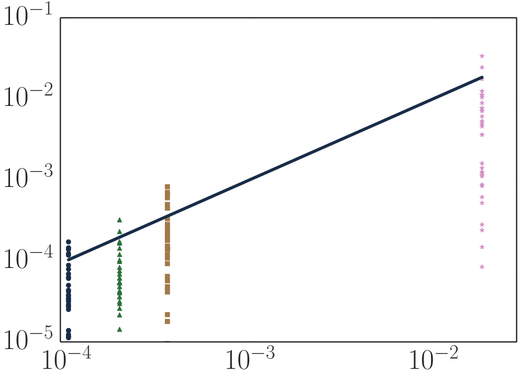
Example: Autocorrelation Function of θ^i, $\pi_1 = 0.2$



Example: Asymptotic Inefficiency InEff_{∞}, $\pi_1 = 0.2$



Example: Small Sample Variance $\mathbb{V}[\bar{h}_N]$ versus HAC Estimates of $\Omega(h)$



Benchmark Random-Walk Metropolis-Hastings (RWMH) Algorithm for DSGE Models

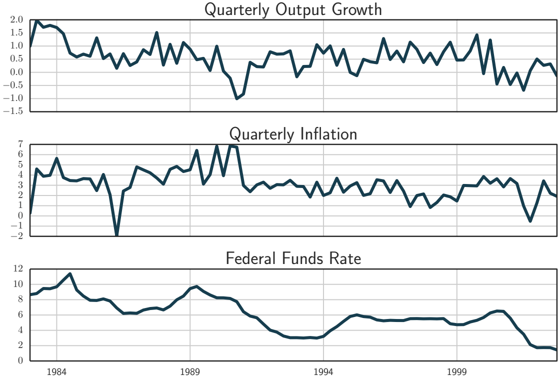
- Initialization:
  1. Use a numerical optimization routine to maximize the log posterior, which up to a constant is given by $\ln p(Y|\theta) + \ln p(\theta)$. Denote the posterior mode by $\hat{\theta}$.
  2. Let $\hat{\Sigma}$ be the inverse of the (negative) Hessian computed at the posterior mode $\hat{\theta}$, which can be computed numerically.
  3. Draw θ^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 ϑ from the proposal distribution $N(\theta^{i-1}, c^2\hat{\Sigma})$.
  2. Set $\theta^i = \vartheta$ with probability

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

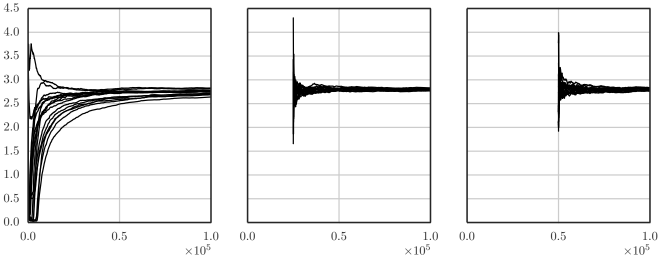
and $\theta^i = \theta^{i-1}$ otherwise.

Observables for Small-Scale New Keynesian Model



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 $\bar{\tau}_{N|N_0}$

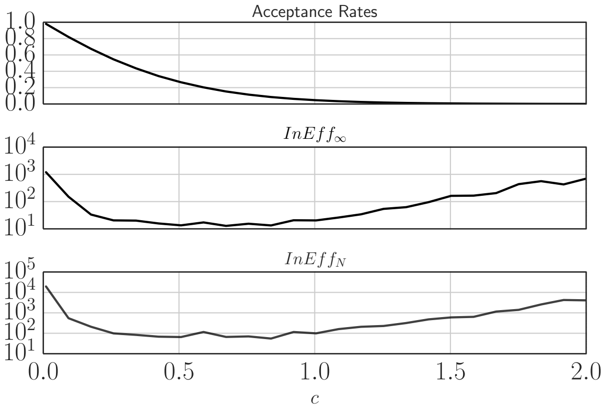


Posterior Estimates of DSGE Model Parameters

Parameter	Mean	[0.05, 0.95]	Parameter	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]			

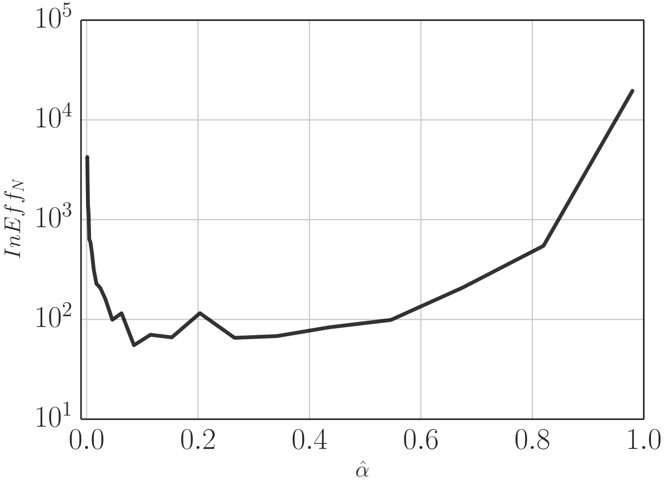
Notes: We generated $N = 100,000$ draws from the posterior and discarded the first 50,000 draws. Based on the remaining draws we approximated the posterior mean and the 5th and 95th percentiles.

DSGE Model Estimation: Effect of Scaling Constant c



Notes: Results are based on $N_{run} = 50$ 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.

DSGE Model Estimation: Acceptance Rate $\hat{\alpha}$ versus Inaccuracy InEff_N

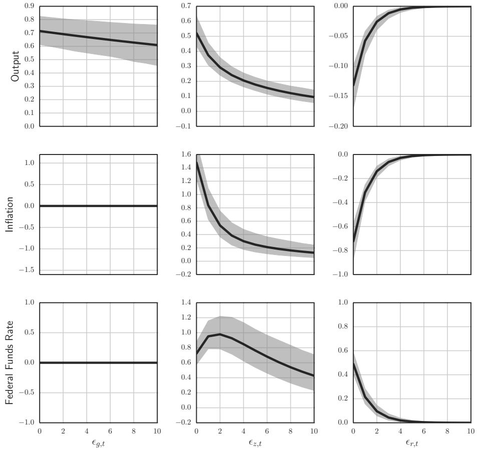


Notes: $\mathrm{InEff}_N[\bar{\tau}]$ versus the acceptance rate $\hat{\alpha}$.

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.

Bayesian Inference – Decision Making

- The posterior expected loss of decision $\delta(\cdot)$:

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

- Bayes decision minimizes the posterior expected loss:

$$\delta^*(Y) = \text{argmin}_d \; \rho\big(\delta(\cdot)|Y\big).$$

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

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

- Then compute

$$\delta_N^*(Y) = \text{argmin}_d \; \bar{\rho}_N\big(\delta(\cdot)|Y\big).$$

Computation of Marginal Data Densities: Modified Harmonic Mean

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

- Geweke (1999):

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

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.

- θ_2 is not identifiable if $\theta_1 = 0$ because θ_2 enters the model only multiplicatively.

- Law of motion of y_t is restricted ARMA(2,1) process:

$$\big(1 - \theta_1^2 L\big)\big(1 - (1 - \theta_1^2)L\big) y_t = \big(1 - \theta_1\theta_2 L\big)\epsilon_t.$$

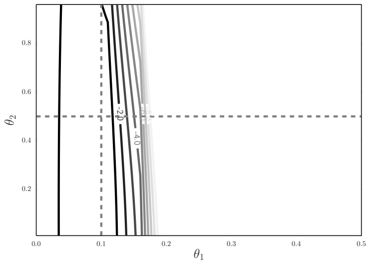
- Given θ_1 and θ_2, we obtain an observationally equivalent process by switching the values of the two roots of the autoregressive lag polynomial.

- Choose $\tilde{\theta}_1$ and $\tilde{\theta}_2$ such that

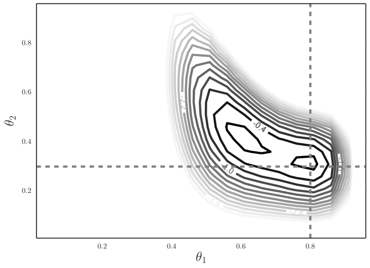
$$\tilde{\theta}_1 = \sqrt{1 - \theta_1^2}, \quad \tilde{\theta}_2 = \theta_1\theta_2/\tilde{\theta}_1.$$

Posteriors for Stylized State-Space Model

Local Identification Problem



Global Identification Problem



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: $\theta_1 = 0.8$, $\theta_2 = 0.3$.

Improvements to MCMC: Blocking

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

$$
\begin{aligned}
& \sqrt{N}(\bar{h}_N - \mathbb{E}[\bar{h}_N]) \\
& \quad \Longrightarrow \quad N\left(0, \frac{1}{N}\sum_{i=1}^{n} \mathbb{V}[h(\theta^i)] + \frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq i} COV\big[h(\theta^i), h(\theta^j)\big]\right).
\end{aligned}
$$

- 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_{blocks} blocks $\theta_1, \ldots, \theta_{N_{blocks}}$ via some rule (perhaps probabilistic), unrelated to the current state of the Markov chain.
2. 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

1. 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, $b = 1, \dots, N_{blocks}$ 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 consists of parameters $(b-1)N_{blocks}, \dots, bN_{blocks}$.[1]
2. Execute Algorithm Block MH Algorithm.

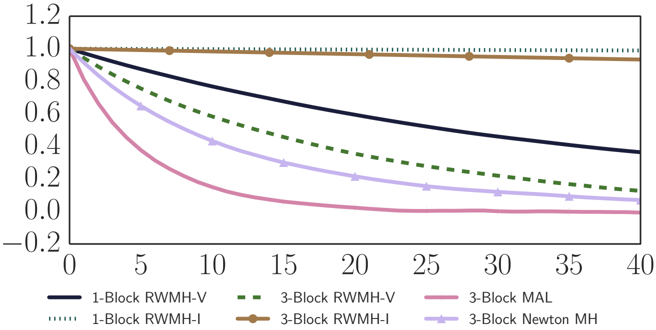
[1] If the number of parameters is not divisible by N_{blocks}, then the size of a subset of the blocks has to be adjusted.

Run Times and Tuning Constants for MH Algorithms

Algorithm	Run Time [hh:mm:ss]	Acceptance 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	$c_1 = 0.400, c_2 = 0.750$
3-Block Newton MH	03:01:40	0.53	$\bar{s} = 0.700, c_2 = 0.600$

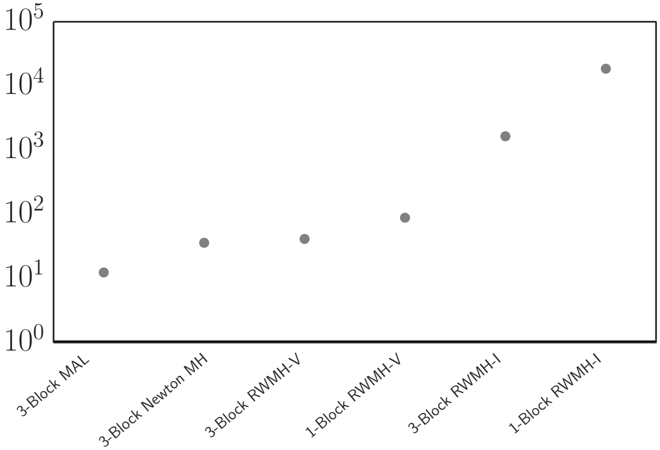
Notes: In each run we generate $N = 100,000$ draws. We report the fastest run time and the average acceptance rate across $N_{run} = 50$ independent Markov chains.
See book for MAL and Newton MH Algorithms.

Autocorrelation Function of τ^i



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

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



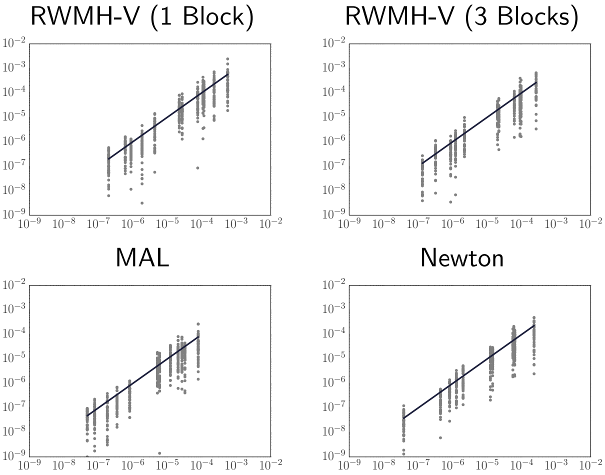
Notes: The small sample inefficiency factors are computed based on $N_{run} = 50$ independent runs of each algorithm.

IID Equivalent Draws Per Second

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

- 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 $\mathbb{V}[\bar{\theta}]$ computed across multiple chains (x-axis) versus the HAC$[\bar{h}]$ estimates of $\Omega(\theta)/N$ (y-axis).