Bayesian Estimation of DSGE Models[1]
Chapter 3: A Crash Course in Bayesian Inference

Ed Herbst[1] Frank Schorfheide[2]

[1]Federal Reserve Board

[2]University of Pennsylvania

February 5, 2016

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Board of Governors or the Federal Reserve System.

Topics

- Deriving a posterior distribution in a linear regression model
- Direct sampling
- Bayesian decision making

Bayesian Inference

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

$$p(\phi|Y) = \frac{p(Y|\phi)p(\phi)}{p(Y)}$$

Linear Regression / AR Models

- Consider AR(1) model:

$$y_t = y_{t-1}\phi + u_t, \quad u_t \sim iidN(0,1).$$

- Let $x_t = y_{t-1}$. Write as

$$y_t = x_t'\phi + u_t, \quad u_t \sim iidN(0,1),$$

or

$$Y = X\phi + U.$$

We can easily allow for multiple regressors. Assume ϕ is $k \times 1$.

- Notice: we treat the variance of the errors as know. The generalization to unknown variance is straightforward but tedious.

- Likelihood function:

$$p(Y|\phi) = (2\pi)^{-T/2} \exp\left\{-\frac{1}{2}(Y - X\phi)'(Y - X\phi)\right\}.$$

A Convenient Prior

- Prior:

$$\phi \sim N\left(0_{k\times 1}, \tau^2 \mathcal{I}_{k\times k}\right), \quad p(\phi) = (2\pi\tau^2)^{-k/2} \exp\left\{-\frac{1}{2\tau^2}\phi'\phi\right\}$$

- Large τ means diffuse prior.
- Small τ means tight prior.

Deriving the Posterior

- Bayes Theorem:

$$\begin{aligned} p(\phi|Y) &\propto p(Y|\phi)p(\phi) \\ &\propto \exp\left\{-\frac{1}{2}[(Y-X\phi)'(Y-X\phi)+\tau^{-2}\phi'\phi]\right\}. \end{aligned}$$

- Guess: what if $\phi|Y \sim N(\bar{\phi}_T, \bar{V}_T)$. Then

$$p(\theta|Y) \propto \exp\left\{-\frac{1}{2}(\phi-\bar{\phi}_T)'\bar{V}_T^{-1}(\phi-\bar{\phi}_T)\right\}.$$

- Rewrite exponential term

$$\begin{aligned} & Y'Y - \phi'X'Y - Y'X\phi + \phi'X'X\phi + \tau^{-2}\phi'\phi \\ &= Y'Y - \phi'X'Y - Y'X\phi + \phi'(X'X+\tau^{-2}\mathcal{I})\phi \\ &= \left(\phi-(X'X+\tau^{-2}\mathcal{I})^{-1}X'Y\right)'\left(X'X+\tau^{-2}\mathcal{I}\right) \\ &\quad \times\left(\phi-(X'X+\tau^{-2}\mathcal{I})^{-1}X'Y\right) \\ &\quad + Y'Y - Y'X(X'X+\tau^{-2}\mathcal{I})^{-1}X'Y. \end{aligned}$$

Deriving the Posterior

- Exponential term is a quadratic function of ϕ.
- Deduce: posterior distribution of ϕ must be a multivariate normal distribution

$$\phi|Y \sim N(\bar{\phi}_T, \bar{V}_T)$$

with

$$\begin{aligned} \bar{\phi}_T &= (X'X + \tau^{-2}\mathcal{I})^{-1}X'Y \\ \bar{V}_T &= (X'X + \tau^{-2}\mathcal{I})^{-1}. \end{aligned}$$

- $\tau \longrightarrow \infty$:

$$\phi|Y \overset{approx}{\sim} N\left(\hat{\phi}_{mle}, (X'X)^{-1}\right).$$

- $\tau \longrightarrow 0$:

$$\phi|Y \overset{approx}{\sim} \text{Pointmass at } 0$$

Marginal Data Density

- Plays an important role in Bayesian model selection and averaging.
- Write

$$\begin{aligned} p(Y) &= \frac{p(Y|\theta)p(\theta)}{p(\theta|Y)} \\ &= \exp\left\{-\frac{1}{2}[Y'Y - Y'X(X'X + \tau^{-2}\mathcal{I})^{-1}X'Y]\right\} \\ &\quad \times (2\pi)^{-T/2}|\mathcal{I} + \tau^2 X'X|^{-1/2}. \end{aligned}$$

- The exponential term measures the goodness-of-fit.
- $|\mathcal{I} + \tau^2 X'X|$ is a penalty for model complexity.

Posterior

- We will often abbreviate posterior distributions $p(\phi|Y)$ by $\pi(\phi)$ and posterior expectations of $h(\phi)$ by

$$\mathbb{E}_\pi[h] = \mathbb{E}_\pi[h(\phi)] = \int h(\phi)\pi(\phi)d\phi = \int h(\phi)p(\phi|Y)d\phi.$$

- We will focus on algorithms that generate draws $\{\phi^i\}_{i=1}^N$ from posterior distributions of parameters in time series models.

- These draws can then be transformed into objects of interest, $h(\phi^i)$, and under suitable conditions a Monte Carlo average of the form

$$\bar{h}_N = \frac{1}{N}\sum_{i=1}^{N} h(\phi^i) \approx \mathbb{E}_\pi[h].$$

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

Direct Sampling

- In the simple linear regression model with Gaussian posterior it is possible to sample directly.
- For $i = 1$ to N, draw ϕ^i from $N\left(\bar{\phi}, \bar{V}_\phi\right)$.
- Provided that $\mathbb{V}_\pi[h(\phi)] < \infty$ we can deduce from Kolmogorov's SLLN and the Lindeberg-Levy CLT that

$$\begin{aligned} \bar{h}_N &\xrightarrow{a.s.} \mathbb{E}_\pi[h] \\ \sqrt{N}\left(\bar{h}_N - \mathbb{E}_\pi[h]\right) &\Longrightarrow N\left(0, \mathbb{V}_\pi[h(\phi)]\right). \end{aligned}$$

Decision Making

- The posterior expected loss associated with a decision $\delta(\cdot)$ is given by

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

- A Bayes decision is a decision that minimizes the posterior expected loss:

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

- Since in most applications it is not feasible to derive the posterior expected risk analytically, we replace $\rho\big(\delta(\cdot)|Y\big)$ by a Monte Carlo approximation of the form

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

- A numerical approximation to the Bayes decision $\delta^*(\cdot)$ is then given by

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

Inference

- Point estimation:
  - Quadratic loss: posterior mean
  - Absolute error loss: posterior median
- Interval/Set estimation $\mathbb{P}_\pi\{\theta \in C(Y)\} = 1 - \alpha$:
  - highest posterior density sets
  - equal-tail-probability intervals

Forecasting

- Example:

$$y_{T+h} = \theta^h y_T + \sum_{s=0}^{h-1} \theta^s u_{T+h-s}$$

- h-step ahead conditional distribution:

$$y_{T+h}|(Y_{1:T}, \theta) \sim N\left(\theta^h y_T, \frac{1-\theta^h}{1-\theta}\right).$$

- Posterior predictive distribution:

$$p(y_{T+h}|Y_{1:T}) = \int p(y_{T+h}|y_T, \theta) p(\theta|Y_{1:T}) d\theta.$$

- For each draw θ^i from the posterior distribution $p(\theta|Y_{1:T})$ sample a sequence of innovations $u^i_{T+1}, \ldots, u^i_{T+h}$ and compute y^i_{T+h} as a function of θ^i, $u^i_{T+1}, \ldots, u^i_{T+h}$, and $Y_{1:T}$.

Model Uncertainty

- Assign prior probabilities $\gamma_{j,0}$ to models M_j, $j = 1, \ldots, J$.
- Posterior model probabilities are given by

$$\gamma_{j,T} = \frac{\gamma_{j,0} p(Y|M_j)}{\sum_{j=1}^{J} \gamma_{j,0} p(Y|M_j)},$$

where

$$p(Y|M_j) = \int p(Y|\theta_{(j)}, M_j) p(\theta_{(j)}|M_j) d\theta_{(j)}$$

- Log marginal data densities are one-step-ahead predictive scores:

$$\begin{aligned} &\ln p(Y|M_j) \\ &= \sum_{t=1}^{T} \ln \int p(y_t|\theta_{(j)}, Y_{1:t-1}, M_j) p(\theta_{(j)}|Y_{1:t-1}, M_j) d\theta_{(j)}. \end{aligned}$$

- Model averaging:

$$p(h|Y) = \sum_{j=1}^{J} \gamma_{j,T} p(h_j(\theta_{(j)})|Y, M_j).$$

A Non-Gaussian Posterior

- Suppose that y_t is determined by the AR(1) model but object of interest is θ, which can be bounded based on ϕ:

$$\phi \leq \theta \quad \text{and} \quad \theta \leq \phi + 1.$$

- Parameter θ is set-identified.
- The interval $\Theta(\phi) = [\phi,\ \phi + 1]$ is called the identified set.
- Prior for θ conditional on ϕ of the form

$$\theta|\phi \sim U[\phi, \phi + 1].$$

A Non-Gaussian Posterior

- Joint posterior of θ and ϕ:

$$p(\theta, \phi|Y) = p(\phi|Y)p(\theta|\phi, Y) \propto p(Y|\phi)p(\theta|\phi)p(\phi).$$

- Since θ does not enter the likelihood function, we deduce that

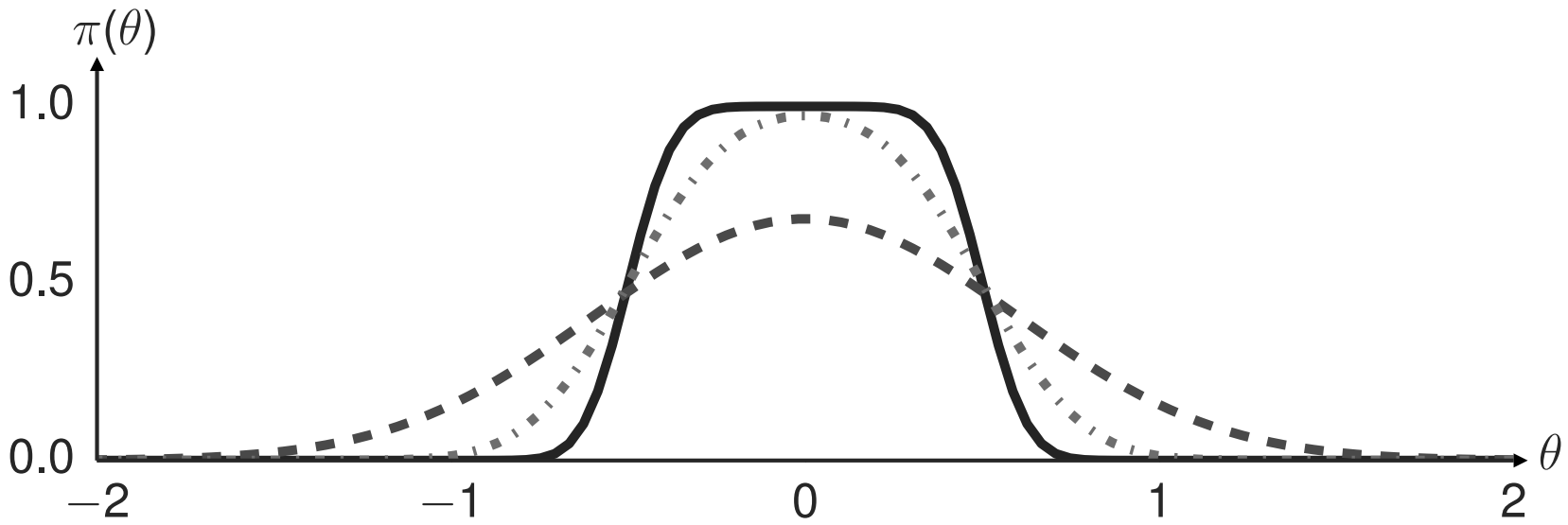
$$\begin{aligned} p(\phi|Y) &= \frac{p(Y|\phi)p(\phi)}{\int p(Y|\phi)p(\phi)d\phi} \\ p(\theta|\phi, Y) &= p(\theta|\phi). \end{aligned}$$

- In our example the marginal posterior distribution of θ is given by

$$\begin{aligned} \pi(\theta) &= \int_{\theta-1}^{\theta} p(\phi|Y)p(\theta|\phi)d\phi \\ &= \Phi_N\left(\frac{\theta - \bar{\phi}}{\sqrt{\bar{V}}}\right) - \Phi_N\left(\frac{\theta - 1 - \bar{\phi}}{\sqrt{\bar{V}}}\right), \end{aligned}$$

where $\Phi_N(x)$ is the cumulative density function of a $N(0, 1)$.

What if the Posterior is Non-Gaussian?



Posterior distribution $\pi(\theta)$ for $\bar{\phi} = -0.5$ and $\bar{V}_\phi$ equal to $1/4$ (dotted), $1/20$ (dashed), and $1/100$ (solid).

Importance Sampling

- Approximate $\pi(\cdot)$ by using a different, tractable density $g(\theta)$ that is easy to sample from.
- For more general problems, posterior density may be non-normalized. So we write

$$\pi(\theta) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{f(\theta)}{Z}.$$

- Importance sampling is based on the identity

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta = \frac{1}{Z}\int_\Theta h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta.$$

- The ratio

$$w(\theta) = \frac{f(\theta)}{g(\theta)}$$

is called the (unnormalized) importance weight.

Importance Sampling

1. For $i = 1$ to N, draw $\theta^i \overset{iid}{\sim} g(\theta)$ and compute the unnormalized importance weights

$$w^i = w(\theta^i) = \frac{f(\theta^i)}{g(\theta^i)}.$$

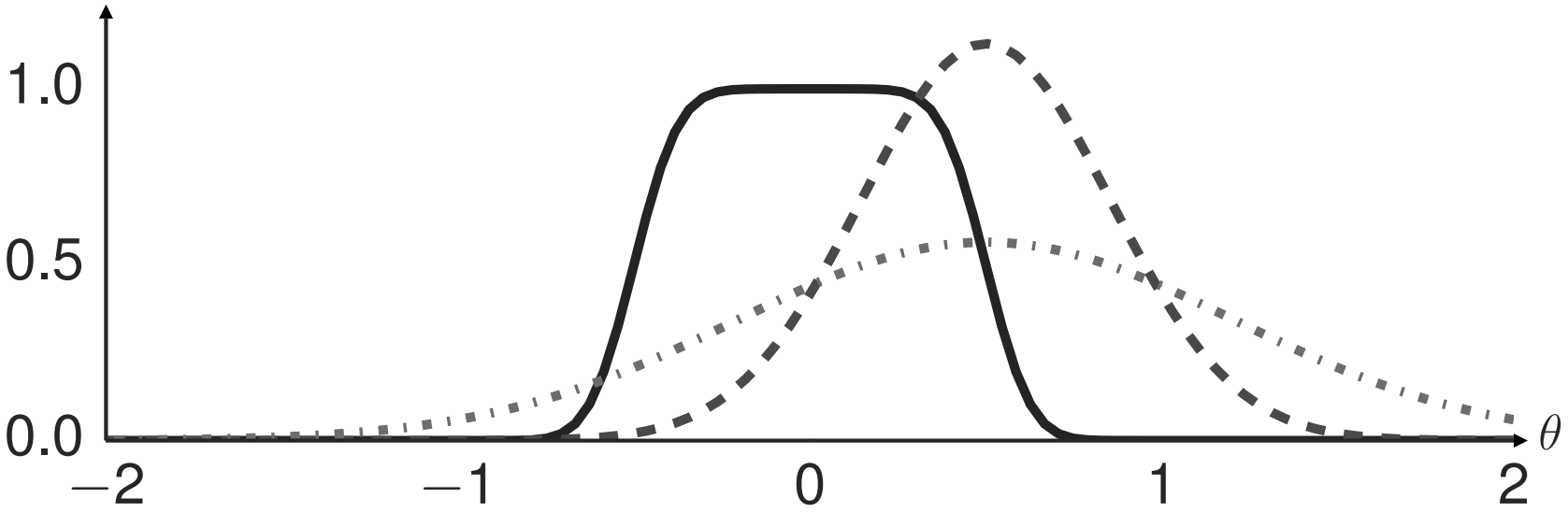
2. Compute the normalized importance weights

$$W^i = \frac{w^i}{\frac{1}{N}\sum_{i=1}^{N} w^i}.$$

An approximation of $\mathbb{E}_\pi[h(\theta)]$ is given by

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

Importance Sampling Distribution



Posterior density $\pi(\theta)$ (solid) as well as two importance sampling densities ("concentrated" (dashed) and "diffuse" (dotted)) $g(\theta)$.

Accuracy

- Since we are generating *iid* draws from $g(\theta)$, it's fairly straightforward to derive a CLT:
- It can be shown that

$$\sqrt{N}(\bar{h}_N - \mathbb{E}_\pi[h]) \Longrightarrow N\big(0, \Omega(h)\big), \quad \text{where} \quad \Omega(h) = \mathbb{V}_g[(\pi/g)(h - \mathbb{E}_\pi[h])].$$

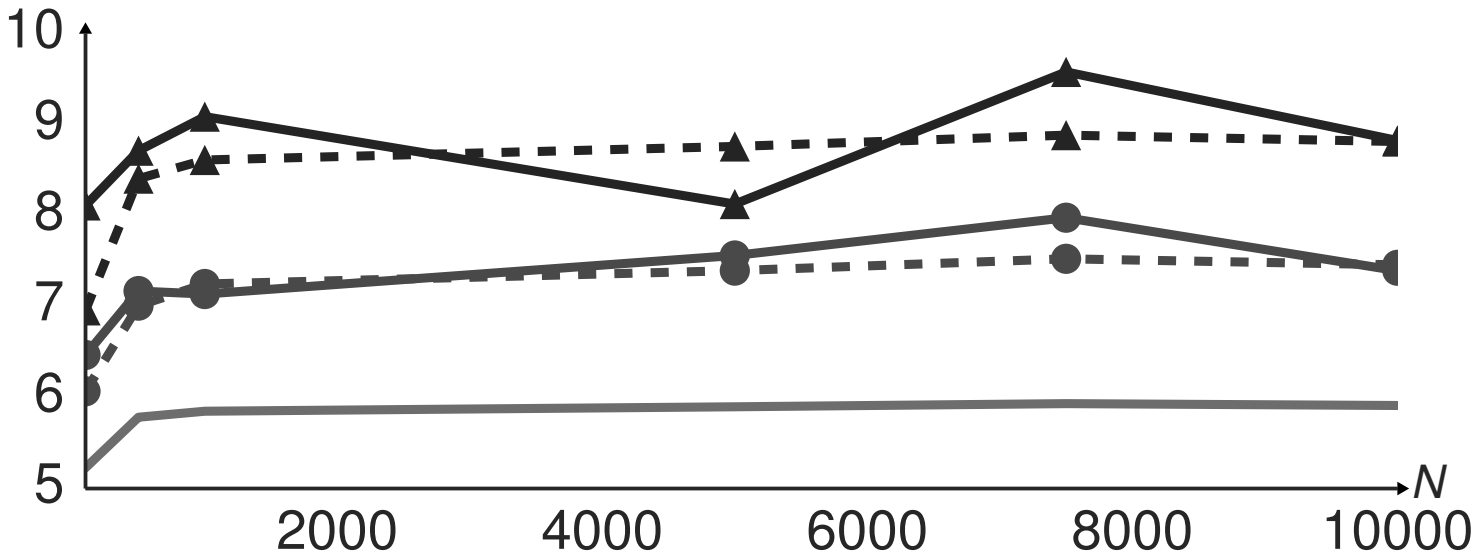
- Using a crude approximation (see, e.g., Liu (2008)), we can factorize $\Omega(h)$ as follows:

$$\Omega(h) \approx \mathbb{V}_\pi[h]\big(\mathbb{V}_g[\pi/g] + 1\big).$$

  The approximation highlights that the larger the variance of the importance weights, the less accurate the Monte Carlo approximation relative to the accuracy that could be achieved with an *iid* sample from the posterior.
- Users often monitor

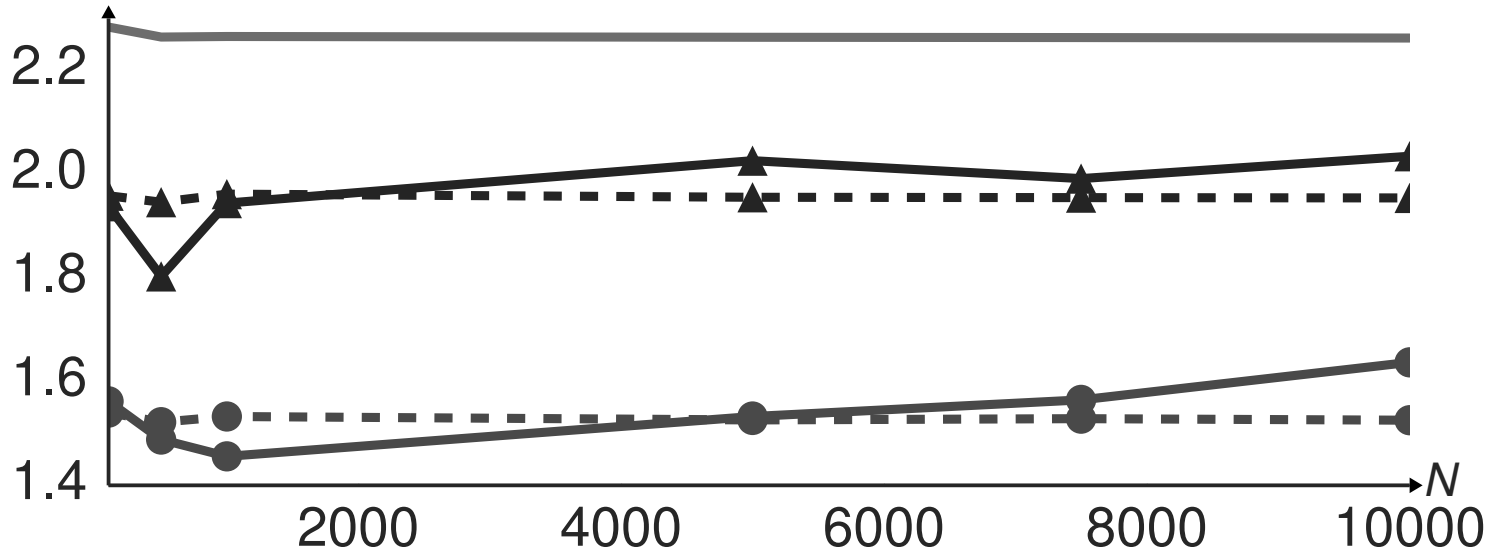
$$ESS = N\frac{\mathbb{V}_\pi[h]}{\Omega(h)} \approx \frac{N}{1 + \mathbb{V}_g[\pi/g]}.$$

Inefficiency Factors for Concentrated IS Density



Large sample inefficiency factors $\text{InEff}_\infty = \Omega(h)/\mathbb{V}_\pi[h]$ (dashed) and as their small sample approximations (solid) based on $N_{run} = 1,000$. We consider $h(\theta) = \theta$ (triangles) and $h(\theta) = \theta^2$ (circles). The solid line (no symbols) depicts the approximate inefficiency factor $1 + \mathbb{V}_g[\pi/g]$.

Inefficiency Factors for Diffuse IS Density



Large sample inefficiency factors $\text{InEff}_\infty = \Omega(h)/\mathbb{V}_\pi[h]$ (dashed) and as their small sample approximations (solid) based on $N_{run} = 1,000$. We consider $h(\theta) = \theta$ (triangles) and $h(\theta) = \theta^2$ (circles). The solid line (no symbols) depicts the approximate inefficiency factor $1 + \mathbb{V}_g[\pi/g]$.

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 $K(\theta|\tilde{\theta})$.

Importance 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:

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

Importance 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:

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

Discrete MH Algorithm

- Iteration i: suppose that $\theta^i = \tau_j$. 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 $\theta^i = \theta^i = \tau_j$.

- Choose (Q terms cancel because of symmetry)

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

Discrete MH Algorithm: 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.

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

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

- 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 $k_{jj}(1 - k_{jj})$.

Convergence

- Autocovariance function of $h(\theta^{(s)})$:

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

- 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).

Convergence

- Define the Monte Carlo estimate

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

- Deduce from CLT

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

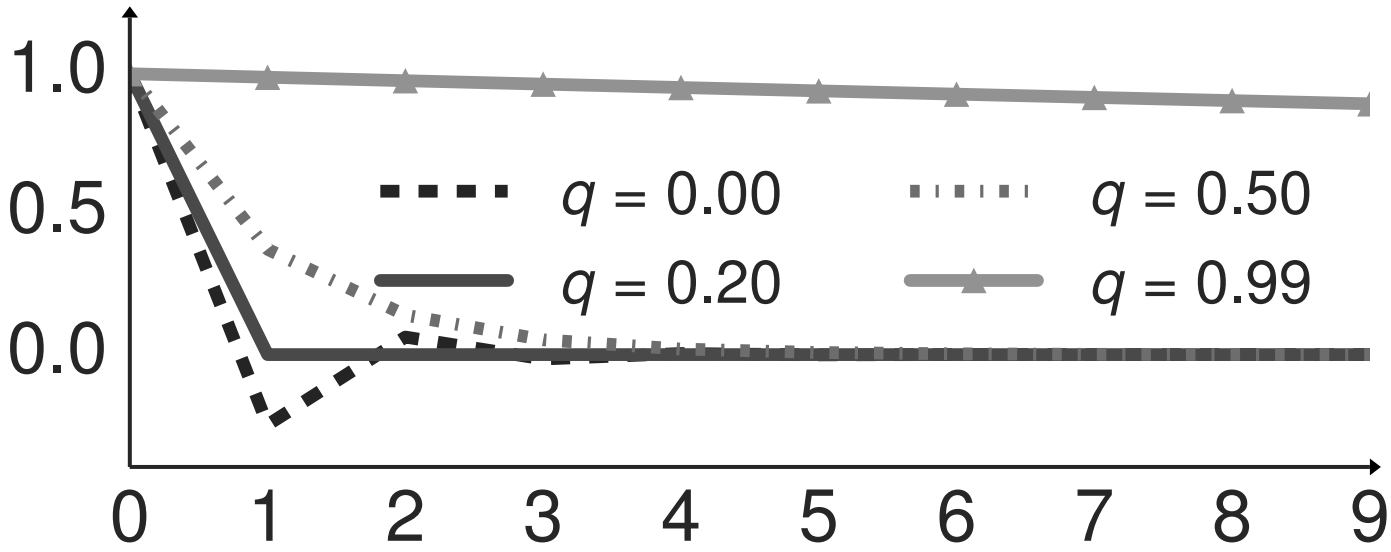
where ΩV_h is the long-run covariance matrix

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

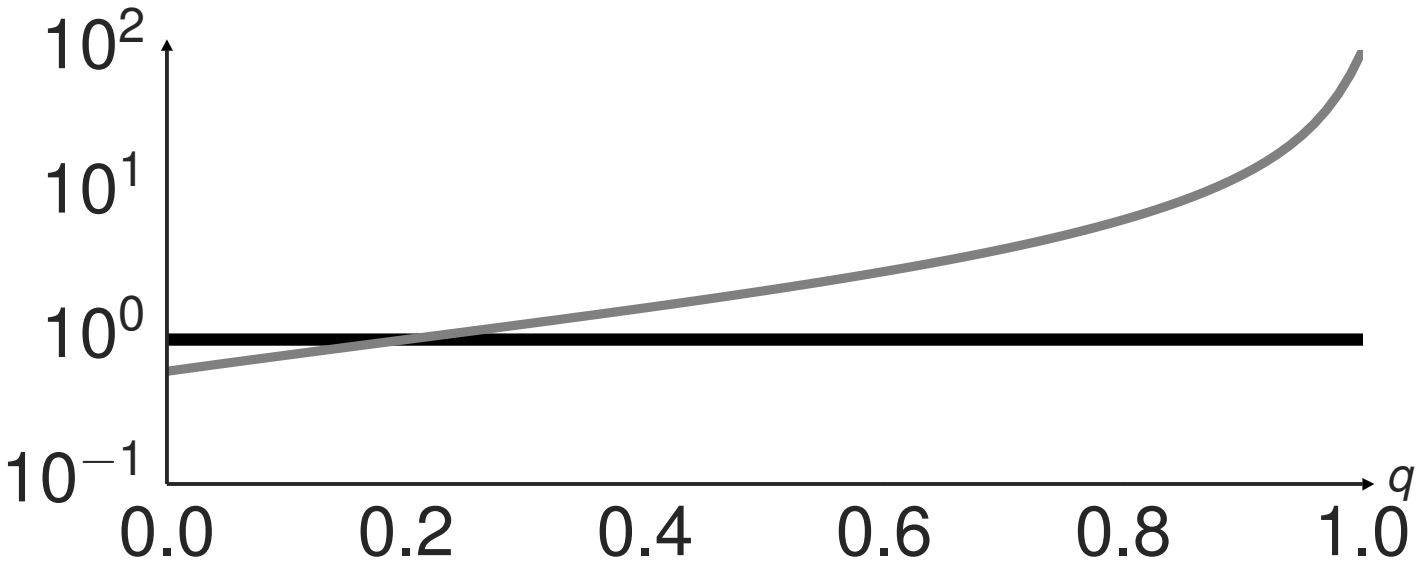
- In turn, the asymptotic inefficiency factor is given by

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

Autocorrelation Function of θ^i



Asymptotic Inefficiency InEff_∞



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