I. Bayesian econometrics

A. Introduction
B. Bayesian inference in the univariate regression model
C. Statistical decision theory
D. Large sample results
E. Diffuse priors
F. Numerical Bayesian methods
    1. Importance sampling

Generic Bayesian problem:

$p(\mathbf{Y}|\boldsymbol{\theta})$ = likelihood (known)

$p(\boldsymbol{\theta})$ = prior (known)

goal: calculate

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{G}$$

for $G = \int p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$

Analytical approach: choose $p(\boldsymbol{\theta})$ from a family such that G can be found with clever algebra.

Numerical approach: satisfied to be able to generate draws

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(D)}$$

from the distribution $p(\boldsymbol{\theta}|\mathbf{Y})$ without ever knowing the distribution (i.e., without calculating G)

Importance sampling:

Step (1): Generate $\boldsymbol{\theta}^{(j)}$ from an (essentially arbitrary) "importance density" $g(\boldsymbol{\theta})$.

Step (2): Calculate

$$\omega^{(j)} = \frac{p[\mathbf{Y}|\boldsymbol{\theta}^{(j)}]p[\boldsymbol{\theta}^{(j)}]}{g[\boldsymbol{\theta}^{(j)}]}.$$

Step (3): Weight the draw $\boldsymbol{\theta}^{(j)}$ by $\omega^{(j)}$ to simulate distribution of $p(\boldsymbol{\theta}|\mathbf{Y})$.

Examples:

$$
\begin{aligned}
E(\boldsymbol{\theta}|\mathbf{Y}) &= \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{Y})\, d\boldsymbol{\theta} \\
&\simeq \frac{\sum_{j=1}^{D} \boldsymbol{\theta}^{(j)} \omega^{(j)}}{\sum_{j=1}^{D} \omega^{(j)}} \\
&\equiv \boldsymbol{\theta}^{*}
\end{aligned}
$$

$$\mathrm{Var}(\boldsymbol{\theta}|\mathbf{Y}) \simeq \frac{\sum_{j=1}^{D}(\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{*})(\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{*})' \omega^{(j)}}{\sum_{j=1}^{D} \omega^{(j)}}$$

$$\mathrm{Prob}(\theta_2 < 0) \simeq \frac{\sum_{j=1}^{D} \delta_{[\theta_2^{(j)} < 0]} \omega^{(j)}}{\sum_{j=1}^{D} \omega^{(j)}}$$

How does this work?

$$\frac{\sum_{j=1}^{D} \boldsymbol{\theta}^{(j)} \omega^{(j)}}{\sum_{j=1}^{D} \omega^{(j)}} = \frac{D^{-1} \sum_{j=1}^{D} \boldsymbol{\theta}^{(j)} \omega^{(j)}}{D^{-1} \sum_{j=1}^{D} \omega^{(j)}}$$

Numerator:

$$D^{-1}\sum_{j=1}^{D}\boldsymbol{\theta}^{(j)}\omega^{(j)} \xrightarrow{p} E[\boldsymbol{\theta}^{(j)}\omega^{(j)}]$$

$$= \int \boldsymbol{\theta}\omega(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$= \int \boldsymbol{\theta}\frac{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}g(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$= \int \boldsymbol{\theta}p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Denominator:

$$D^{-1}\sum_{j=1}^{D}\omega^{(j)} \xrightarrow{p} E[\omega^{(j)}]$$

$$= \int \omega(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$= \int \frac{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}g(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$= \int p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$= p(\mathbf{Y})$$

Conclusion:

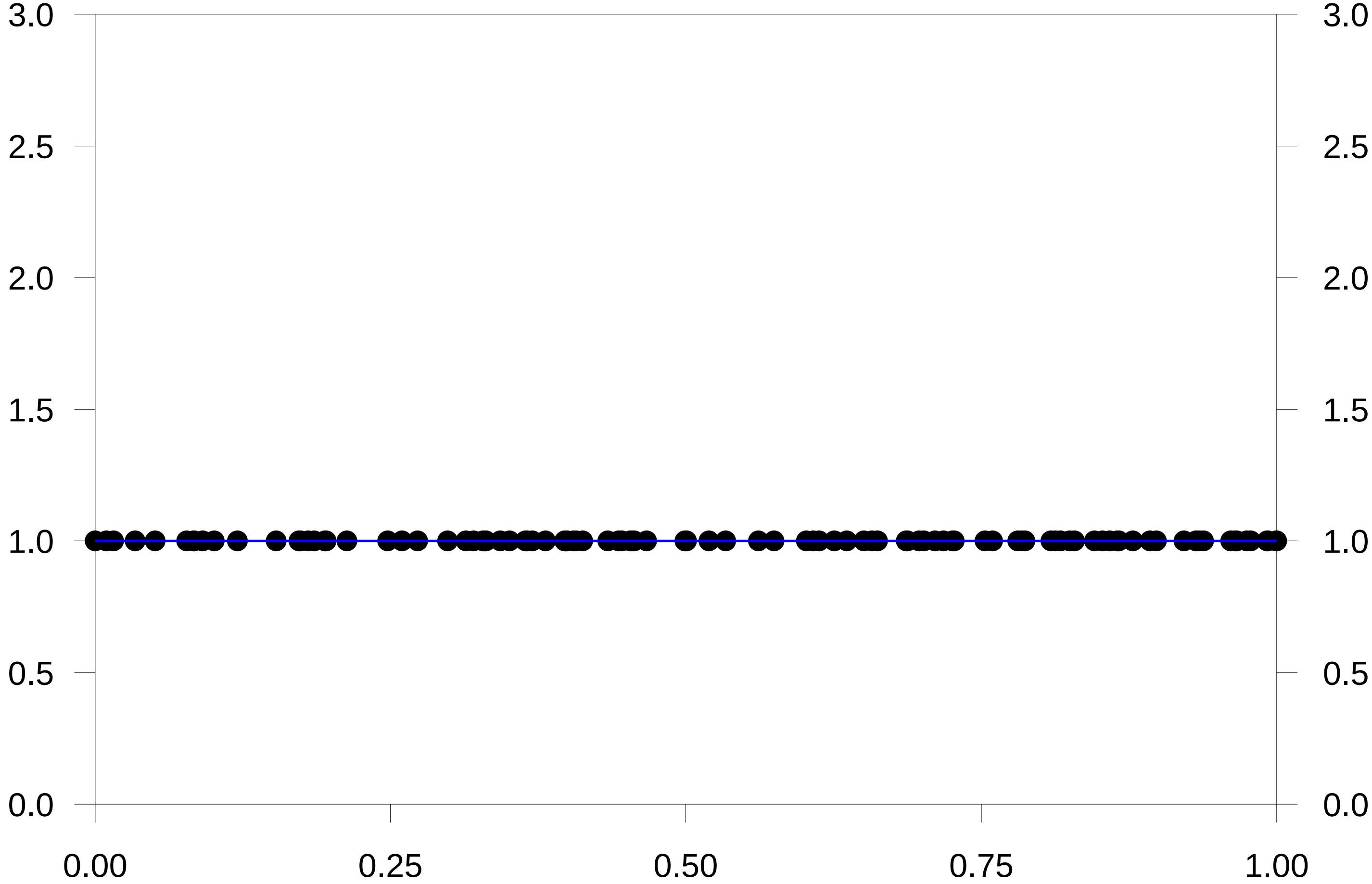
$$\frac{\sum_{j=1}^{D} \boldsymbol{\theta}^{(j)} \omega^{(j)}}{\sum_{j=1}^{D} \omega^{(j)}} \xrightarrow{p} \frac{\int \boldsymbol{\theta} p(\mathbf{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{p(\mathbf{Y})}$$

$$= \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta}$$

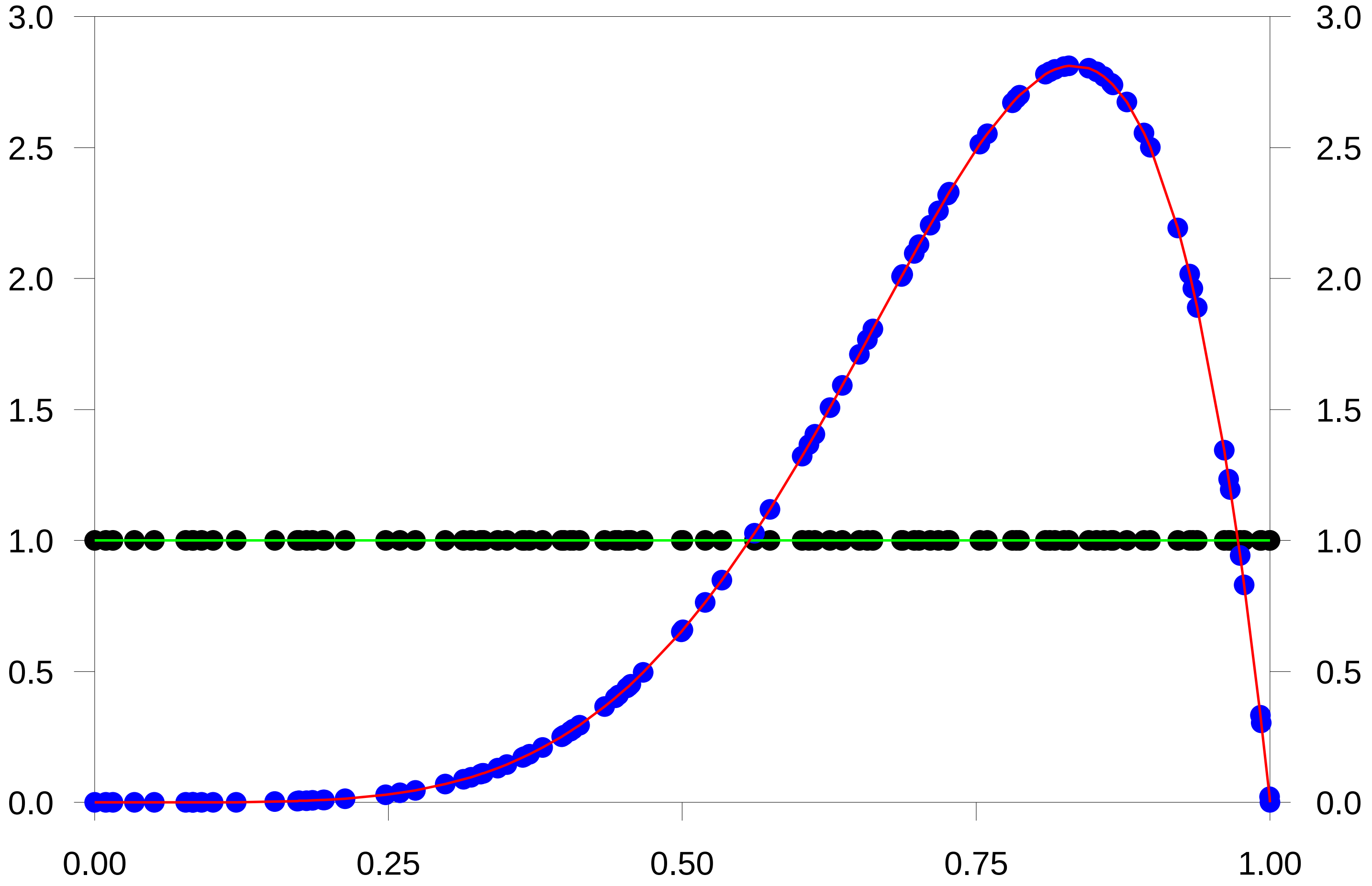
Example: $\theta \in [0,1]$

Importance density $g(\theta)$: $\theta \sim U(0,1)$

Draws from uniform importance density

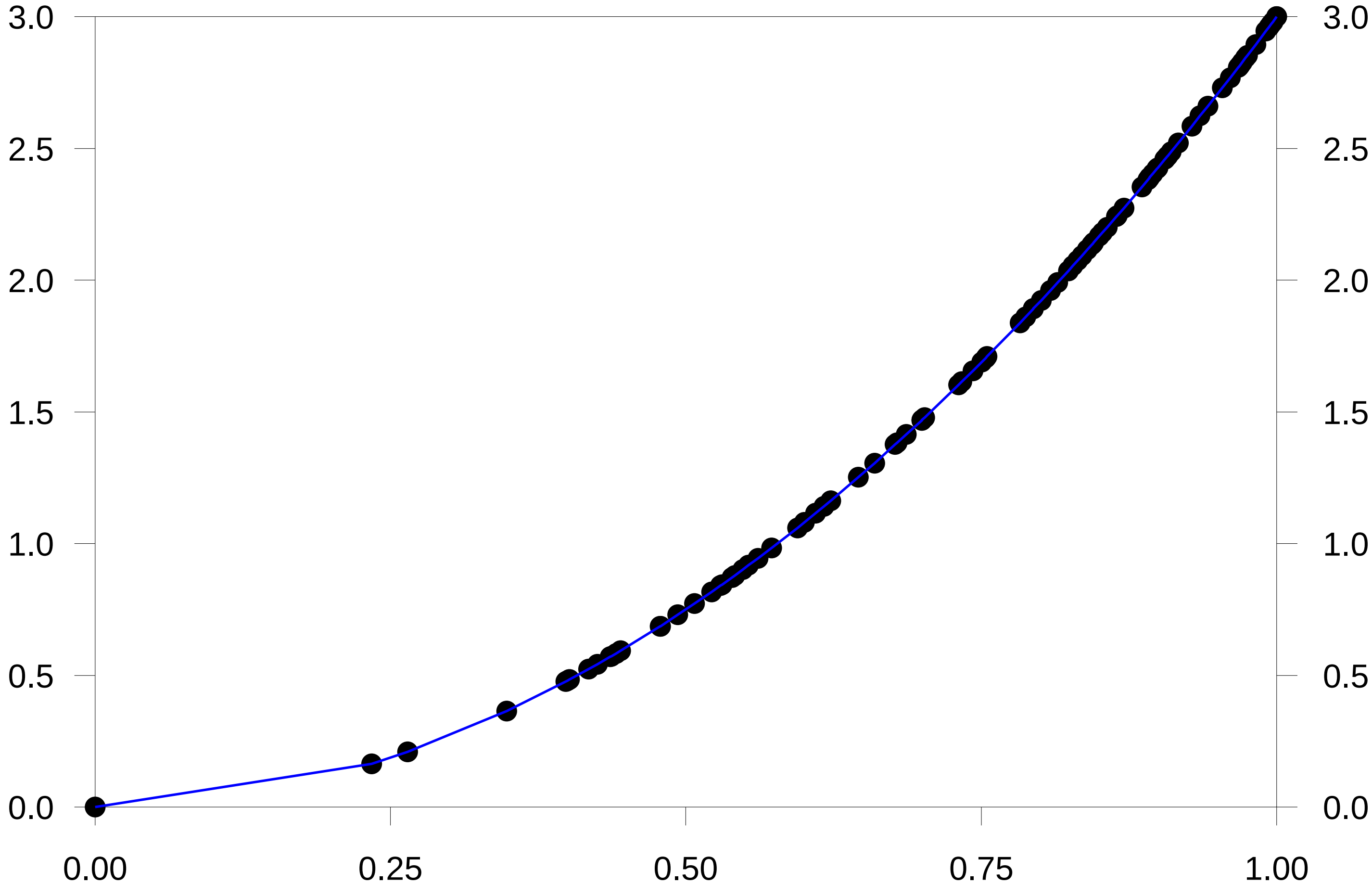


Reweighted draws

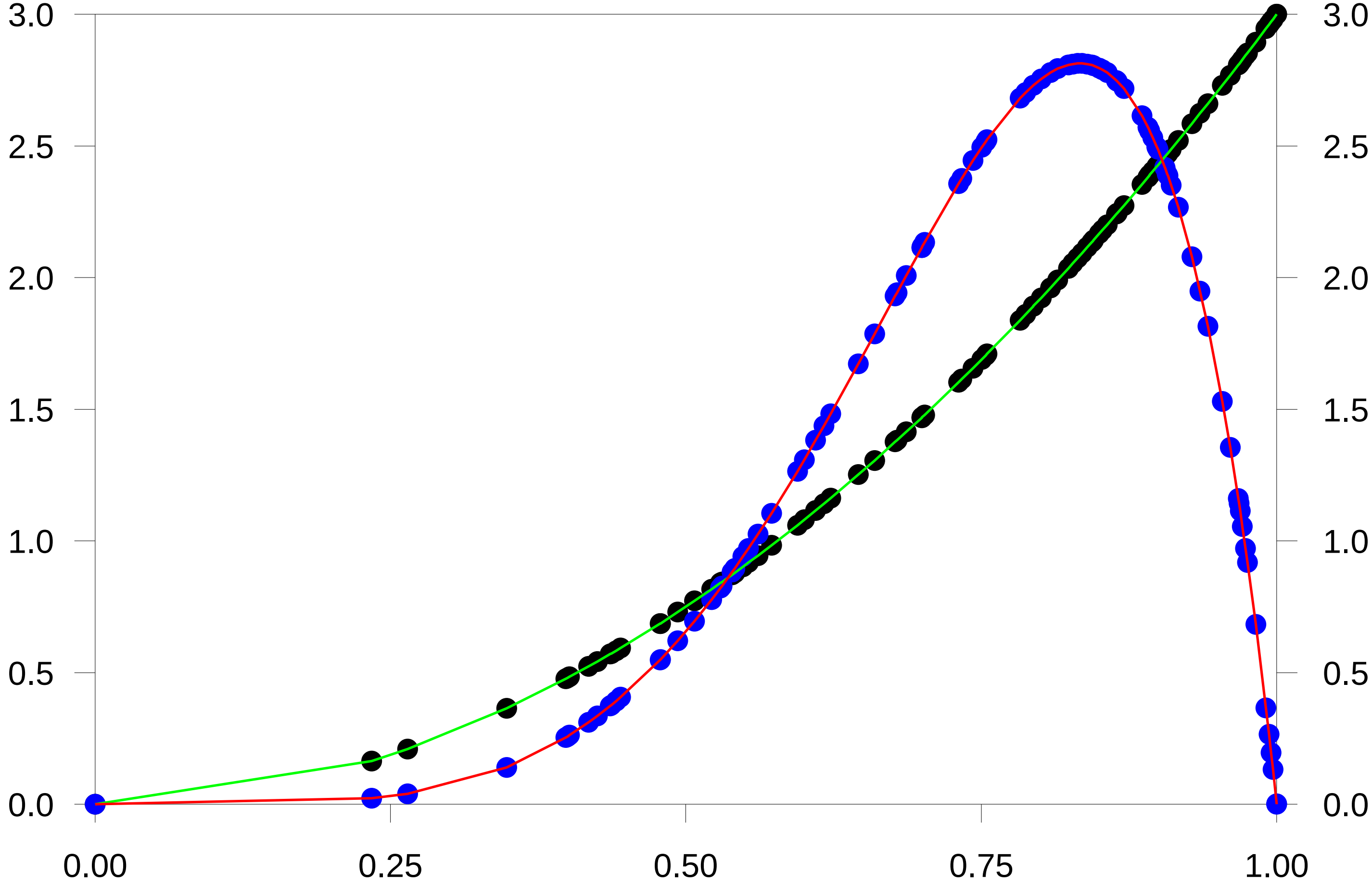


- Algorithm will converge faster the more the importance density resembles the target

Draws from the importance density $g(x) = 3x^2$



Reweighted draws



What's required of $g(.)$?

$$\boldsymbol{\theta}^{(j)}\omega^{(j)} = \frac{\boldsymbol{\theta}^{(j)} p[\mathbf{Y}|\boldsymbol{\theta}^{(j)}] p[\boldsymbol{\theta}^{(j)}]}{g[\boldsymbol{\theta}^{(j)}]}$$ should

satisfy Law of Large Numbers.

Khintchine's Theorem: If $\{\mathbf{x}_j\}_{j=1}^{D}$ is i.i.d.

with finite mean $\boldsymbol{\mu}$, then $D^{-1}\sum_{j=1}^{D}\mathbf{x}_j \xrightarrow{p} \boldsymbol{\mu}$

Note:

- does not require $\mathbf{x}_j$ to have

finite variance

- $\boldsymbol{\theta}^{(j)}$ are drawn i.i.d. from $g(\boldsymbol{\theta})$

by construction

So we only need

$$E(\boldsymbol{\theta}|\mathbf{Y}) = \int_{\aleph} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta}$$ exists

$$p(\boldsymbol{\theta}|\mathbf{Y}) = k p(\mathbf{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

support of $g(\boldsymbol{\theta})$ includes $\aleph$

However, convergence may be very slow if variance of

$$\frac{\boldsymbol{\theta}^{(j)} p[\mathbf{Y}|\boldsymbol{\theta}^{(j)}] p[\boldsymbol{\theta}^{(j)}]}{g[\boldsymbol{\theta}^{(j)}]}$$

is infinite.

Practical observations:

- works best if $g(\boldsymbol{\theta})$ has fatter tails than $p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
- works best when $g(\boldsymbol{\theta})$ is good approximation to $p(\boldsymbol{\theta}|\mathbf{Y})$

Always produces an answer, good idea to check it.

(1) Try special cases where result is known analytically.

(2) Try different $g(.)$ to see if get the same result.

(3) Use analytic results for components of θ in order to keep dimension that must be importance-sampled small.

I. Bayesian econometrics

F. Numerical Bayesian methods

1. Importance sampling
2. The Gibbs sampler

Suppose the parameter vector $\boldsymbol{\theta}$ can be partitioned as $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_3')$ with the property that $p(\boldsymbol{\theta}|\mathbf{Y})$ is of unknown form but

$$p(\boldsymbol{\theta}_1|\mathbf{Y}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$$

$$p(\boldsymbol{\theta}_2|\mathbf{Y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3)$$

$$p(\boldsymbol{\theta}_3|\mathbf{Y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

are of known form (same idea works for 2, 4, or n blocks)

(1) Start with arbitrary initial guesses

$\boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)}$ for $j = 1$.

(2) Generate:

$\boldsymbol{\theta}_1^{(j+1)}$ from $p(\boldsymbol{\theta}_1 | \mathbf{Y}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)})$

$\boldsymbol{\theta}_2^{(j+1)}$ from $p(\boldsymbol{\theta}_2 | \mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)})$

$\boldsymbol{\theta}_3^{(j+1)}$ from $p(\boldsymbol{\theta}_3 | \mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_2^{(j+1)})$

(3) Repeat step (2) for $j = 1, 2, \ldots, D$

Notice the sequence $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{D}$ is a Markov chain with transition kernel

$$\pi(\boldsymbol{\theta}^{(j+1)} | \boldsymbol{\theta}^{(j)}) = p(\boldsymbol{\theta}_3^{(j+1)} | \mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_2^{(j+1)})$$

$$p(\boldsymbol{\theta}_2^{(j+1)} | \mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)})$$

$$p(\boldsymbol{\theta}_1^{(j+1)} | \mathbf{Y}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)})$$

Under quite general conditions, the realizations from a Markov chain for $D \to \infty$ converge to draws from the ergodic distribution of the chain $\pi(\boldsymbol{\theta})$ satisfying

$$\pi(\boldsymbol{\theta}^{(j+1)}) = \int_{\Re^k} \pi(\boldsymbol{\theta}^{(j+1)} | \boldsymbol{\theta}^{(j)}) \pi(\boldsymbol{\theta}^{(j)}) d\boldsymbol{\theta}^{(j)}$$

Claim: the ergodic distribution of this chain corresponds to the posterior distribution:

$$\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{Y})$$

Proof:

$$\int_{\Re^k} \pi(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)})\pi(\boldsymbol{\theta}^{(j)})d\boldsymbol{\theta}^{(j)}$$

$$= \int_{\Re^k} \Big\{ p(\boldsymbol{\theta}_3^{(j+1)}|\mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_2^{(j+1)})$$

$$p(\boldsymbol{\theta}_2^{(j+1)}|\mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)})$$

$$p(\boldsymbol{\theta}_1^{(j+1)}|\mathbf{Y}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)}) \Big\}$$

$$p(\boldsymbol{\theta}^{(j)}|\mathbf{Y})d\boldsymbol{\theta}^{(j)}$$

$$= \int_{\Re^k} p(\boldsymbol{\theta}^{(j+1)}, \boldsymbol{\theta}^{(j)} | \mathbf{Y}) d\boldsymbol{\theta}^{(j)}$$

$$= p(\boldsymbol{\theta}^{(j+1)} | \mathbf{Y})$$

Implication: if we throw out the first D_0 draws (for D_0 large), then $\boldsymbol{\theta}^{(D_0+1)}, \boldsymbol{\theta}^{(D_0+2)}, \ldots, \boldsymbol{\theta}^{(D)}$ represent draws from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$.

Checks:

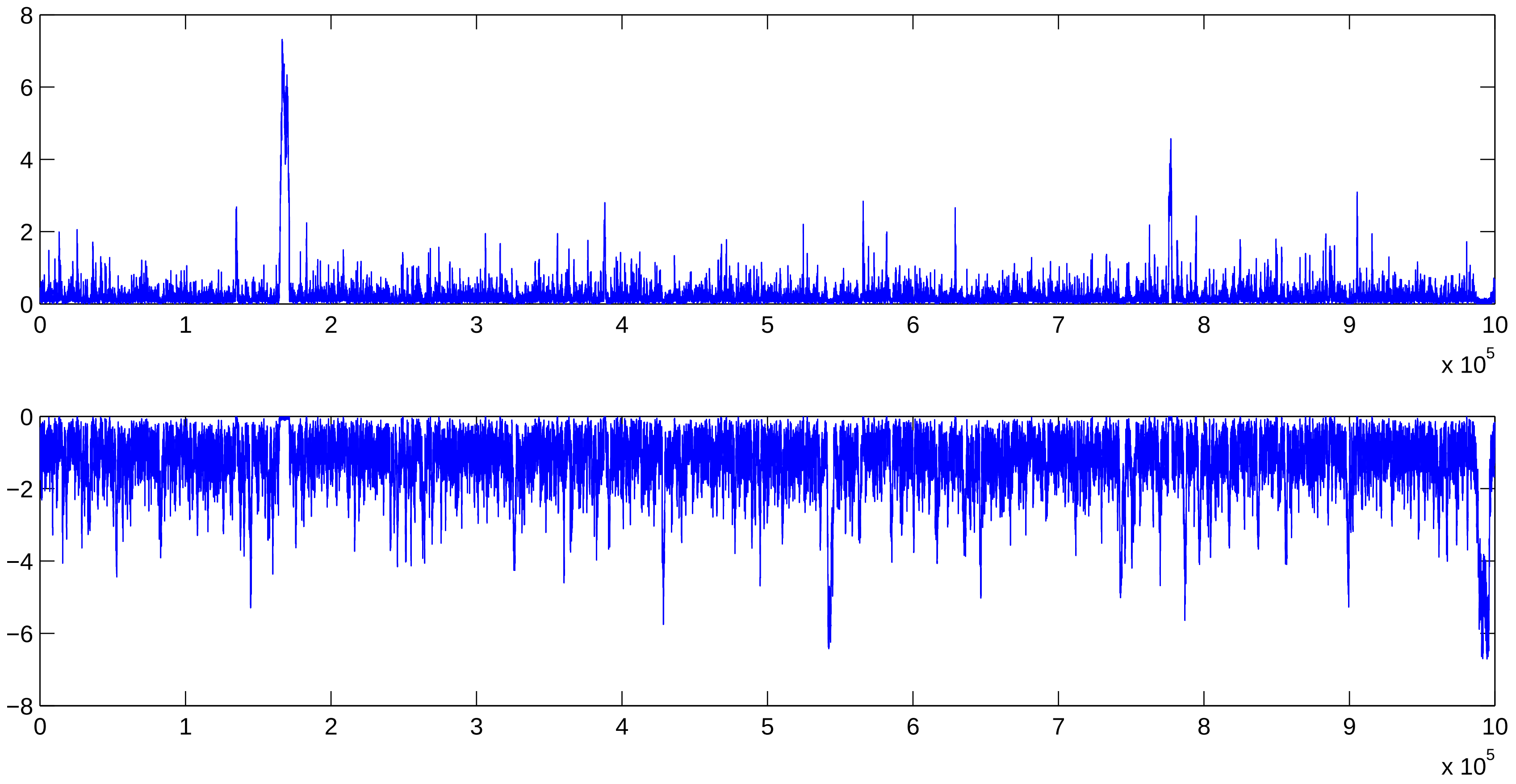
(1) Change $\theta^{(1)} \Rightarrow$ same answer?

(2) Change $D_0, D \Rightarrow$ same answer?

(3) Plot elements of $\theta^{(j)}$ as function of j

to see if it looks same across blocks.

Example of bad mixing



Checks:

(4) Calculate autocorrelations of elements of $\theta^{(j)}$

– Note: throwing out observations does not "cure" the problem

(5) Do formal statistical tests for stability

Geweke's diagnostic:

Test whether mean of θ_i for first 10% of draws is same as for last 50%.

Repeat for each parameter i.

(1) Calculate mean of parameter i over first subsample:

$$q_1 = N_1^{-1} \sum_{j=1}^{N_1} \theta_i^{(j)}$$

for say $N_1 = 0.1D$

(2) Estimate $\hat{s}_1 = 2\pi$ times spectrum at frequency 0 over this subsample

(3) Do same for second subsample, e.g.

$$q_2 = N_2^{-1} \sum_{j=J_2+1}^{D} \theta_i^{(j)}$$

for $J_2 = N_2 = 0.5D$

(4) Calculate

$$\frac{q_1 - q_2}{\sqrt{\hat{s}_1/N_1 + \hat{s}_2/N_2}} \xrightarrow{d} N(0,1),$$

e.g., reject stability if exceeds ± 1.96

Popular approach to estimate s

(e.g., Dynare):

(a) Divide subsample into 100 blocks

(i.e., block 1 $=$ first 1% of draws)

(b) Calculate mean over each block and

autocovariances of these means

(c) Use Newey-West with 4, 8, or 15 lags

($=$ 4%, 8%, or 15% of sample) to get $\hat{s}_1$

I. Bayesian econometrics

F. Numerical Bayesian methods

1. Importance sampling
2. The Gibbs sampler
3. Metropolis-Hastings algorithm

Suppose $\{s_t\}_{t=1}^{T}$ is an ergodic K-state Markov chain,

$$s_t \in \{1, 2, \ldots, K\}$$

with transition probabilities

$$p_{ij} = \Pr[s_t = j | s_{t-1} = i]$$

$$\sum_{j=1}^{K} p_{ij} = 1 \quad \text{for } i = 1, \ldots, K$$

$$p_{ij} \geq 0 \quad \text{for } i, j = 1, \ldots, K$$

The ergodic or unconditional probabilities satisfy

$$\Pr[s_t = j] = \sum_{i=1}^{K} \Pr[s_t = j, s_{t-1} = i]$$

$$\pi_j = \sum_{i=1}^{K} p_{ij}\pi_i$$

Proposition: Suppose we can find a set of numbers $f_1, f_2, \ldots, f_K$ such that

$$f_j \geq 0 \quad \text{for } j = 1, \ldots, K$$

$$\sum_{j=1}^{K} f_j = 1$$

$$f_i p_{ij} = f_j p_{ji}$$

Then $f_j = \pi_j$

Proof: We're given that

$$f_i p_{ij} = f_j p_{ji}$$

sum over i:

$$\sum_{i=1}^{K} f_i p_{ij} = f_j \sum_{i=1}^{K} p_{ji} = f_j$$

which satisfy definitions of π_i,

$$\sum_{i=1}^{K} \pi_i p_{ij} = \pi_j$$

Works also for continuous-valued Markov chains.

If $\mathbf{x}_t \in \Re^k$ is Markov with transition kernel $p(\mathbf{x}, \mathbf{y})$ (meaning that):

$$\Pr[\mathbf{x}_t \in A | \mathbf{x}_{t-1} = \mathbf{x}] = \int_A p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$$

then the ergodic density $\pi(\mathbf{y})$, which signifies that

$$\Pr[\mathbf{x}_t \in A] = \int_A \pi(\mathbf{y}) d\mathbf{y},$$

satisfies

$$\pi(\mathbf{y}) = \int_{\Re^k} p(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x}) d\mathbf{x}$$

Proposition: if

$f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$

$\int_{\Re^k} f(\mathbf{x}) d\mathbf{x} = 1$

$f(\mathbf{x}) p(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}) p(\mathbf{y}, \mathbf{x})$

for all $\mathbf{x}, \mathbf{y}$

then

$\pi(\mathbf{x}) = f(\mathbf{x})$

Goal in Metropolis-Hastings:

We know how to calculate $h\pi(\mathbf{x})$ (where h may be an unknown constant) and want to sample from it.

Solution: generate a sample $\{\mathbf{x}_t\}$ from a Markov chain whose ergodic density is $\pi(\mathbf{x})$

How MH works:

We previously generated $\mathbf{x}_{t-1} = \mathbf{x}$

We now generate a candidate

$\mathbf{y}$ from some known density $q(\mathbf{x}, \mathbf{y})$

We'll then set $\mathbf{x}_t = \mathbf{y}$ if $\pi(\mathbf{y})/\pi(\mathbf{x})$

is big and otherwise keep $\mathbf{x}_t = \mathbf{x}$

Let $\alpha(\mathbf{x}, \mathbf{y})$ be probability we set

$$\mathbf{x}_t = \mathbf{y}$$

If $\pi(\mathbf{x})q(\mathbf{x},\mathbf{y}) > 0$, then

$$\alpha(\mathbf{x},\mathbf{y}) = \min\left[\frac{\pi(\mathbf{y})q(\mathbf{y},\mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x},\mathbf{y})}, 1\right]$$

otherwise

$$\alpha(\mathbf{x},\mathbf{y}) = 1$$

When $\mathbf{x} \neq \mathbf{y}$, the transition kernel of this chain is $q(\mathbf{x}, \mathbf{y})\alpha(\mathbf{x}, \mathbf{y})$. To show that $\pi(\mathbf{y})$ is the ergodic density of this chain, we must show that

$$\pi(\mathbf{x})\alpha(\mathbf{x},\mathbf{y})q(\mathbf{x},\mathbf{y})$$
$$= \pi(\mathbf{y})\alpha(\mathbf{y},\mathbf{x})q(\mathbf{y},\mathbf{x})$$

But

$$\pi(\mathbf{x})\alpha(\mathbf{x},\mathbf{y})q(\mathbf{x},\mathbf{y})$$
$$= \pi(\mathbf{x})q(\mathbf{x},\mathbf{y})\min\left[\frac{\pi(\mathbf{y})q(\mathbf{y},\mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x},\mathbf{y})},1\right]$$
$$= \min[\pi(\mathbf{y})q(\mathbf{y},\mathbf{x}),\pi(\mathbf{x})q(\mathbf{x},\mathbf{y})]$$
$$= \pi(\mathbf{y})\alpha(\mathbf{y},\mathbf{x})q(\mathbf{y},\mathbf{x})$$

Options for candidate density:

(1) independent $q(\mathbf{y}, \mathbf{x}) = q(\mathbf{y})$

e.g., $\mathbf{y} \sim N(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ where $\boldsymbol{\lambda}$ is our guess of mean of $\pi(\mathbf{y})$

Options for candidate density:

(2) random walk

$$q(\mathbf{y},\mathbf{x}) = q(\mathbf{y}-\mathbf{x})$$

e.g.,

$$q(\mathbf{y},\mathbf{x}) = (2\pi)^{-n/2}|\mathbf{\Lambda}|^{-1/2} \times \exp[(-1/2)(\mathbf{y}-\mathbf{x})'\mathbf{\Lambda}^{-1}(\mathbf{y}-\mathbf{x})]$$