Bayesian analysis of VARs, State-space models and DSGEs
Part III - Estimation by Simulation

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- **Monte Carlo**
- **Gibbs sampling**
- **Markov Chain Monte Carlo (MCMC) simulation**

Monte Carlo sampling

- If $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ is an *iid* **sequence** from a distribution $p(\theta)$, then

$$\frac{1}{N}\sum_{t=1}^{N}\theta^{(t)} \quad\rightarrow\quad E(\theta)$$

$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) \quad\rightarrow\quad E[g(\theta)]$$

  where $g(\theta)$ is some well-behaved function.

- Easy to compute **tail probabilities** $\Pr(\theta \leq c)$ by letting

$$g(\theta) = I\left(\theta \leq c\right)$$

  and

$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) = \frac{\#\ \theta\text{-draws smaller than } c}{N}.$$

Gibbs sampling

- Easily implemented methods for **sampling from multivariate distributions**, $p(\theta_1, ..., \theta_k)$.
- Requirements: Easily sampled **full conditional posteriors**:
  - $p(\theta_1 | \theta_2, \theta_3 ..., \theta_k)$
  - $p(\theta_2 | \theta_1, \theta_3, ..., \theta_k)$
  - $\vdots$
  - $p(\theta_k | \theta_1, \theta_2, ..., \theta_{k-1})$
- Started out in the early 80's in the image analysis literature.
- Gibbs sampling is a **special case of Metropolis-Hastings (MCMC)**

The Gibbs sampling algorithm

A: Choose initial values $\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_n^{(0)}$.

B: B_1 Draw $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_n^{(0)})$

B_2 Draw $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_n^{(0)})$

$\vdots$

B_n Draw $\theta_n^{(1)}$ from $p(\theta_n|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{n-1}^{(1)})$

C: Repeat Step B N times.

Gibbs sampling, cont.

- The Gibbs draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are **dependent** (autocorrelated), but **arithmetic means converge to expected values**

$$\frac{1}{N}\sum_{t=1}^{N}\theta_j^{(t)} \quad \rightarrow \quad E(\theta_j)$$

$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) \quad \rightarrow \quad E[g(\theta)]$$

- $\theta^{(1)}, ..., \theta^{(N)}$ **converges in distribution** to the target $p(\theta)$.
- $\theta_j^{(1)}, ..., \theta_j^{(N)}$ converge to the marginal distribution of θ_j, $p(\theta_j)$.
- **Dependent** draws $\rightarrow$ **less efficient** than iid sampling.
- Compare sampling from:
  - $x_t \overset{iid}{\sim} N(0, \sigma^2)$
  - $x_t = 0.9x_{t-1} + \varepsilon_t$ with $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$.

GIBBS SAMPLING MULTIVARIATE NORMAL

- Bivariate normal:
  - Joint distribution

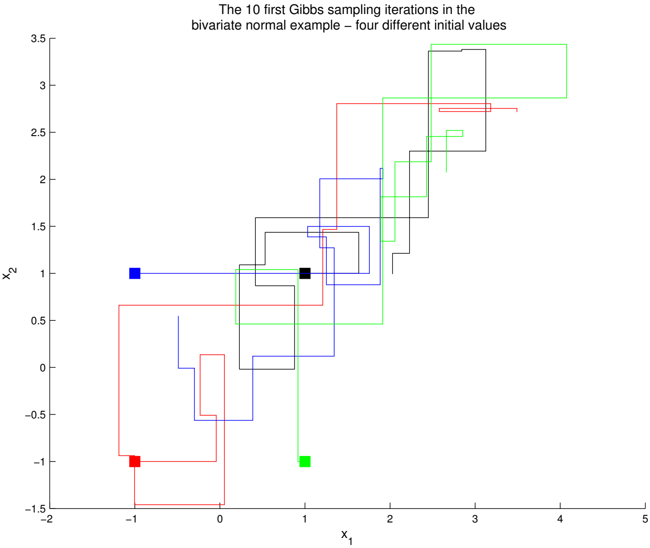
$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

  - Full conditional posteriors:

$$\begin{aligned} \theta_1 | \theta_2 &\sim N[\mu_1 + \rho(\theta_2 - \mu_2), 1 - \rho^2] \\ \theta_2 | \theta_1 &\sim N[\mu_2 + \rho(\theta_1 - \mu_1), 1 - \rho^2] \end{aligned}$$

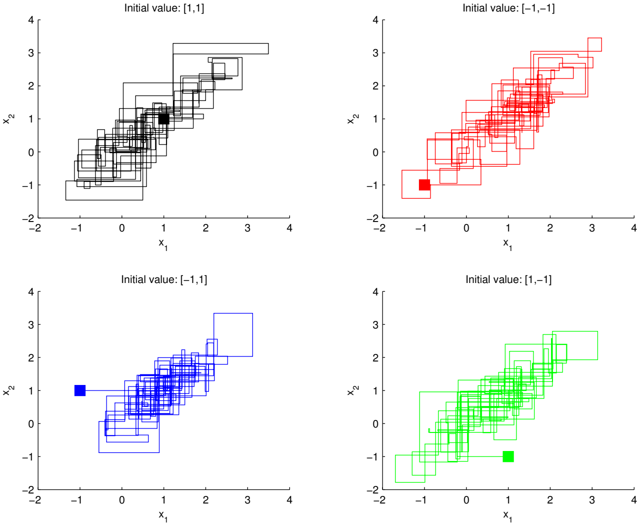
VOID

Gibbs sampling - Bivariate normal

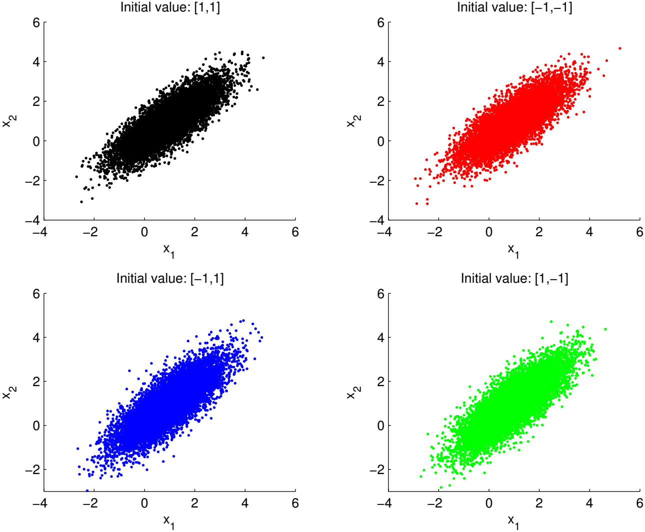


VOID

Gibbs sampling - Bivariate normal

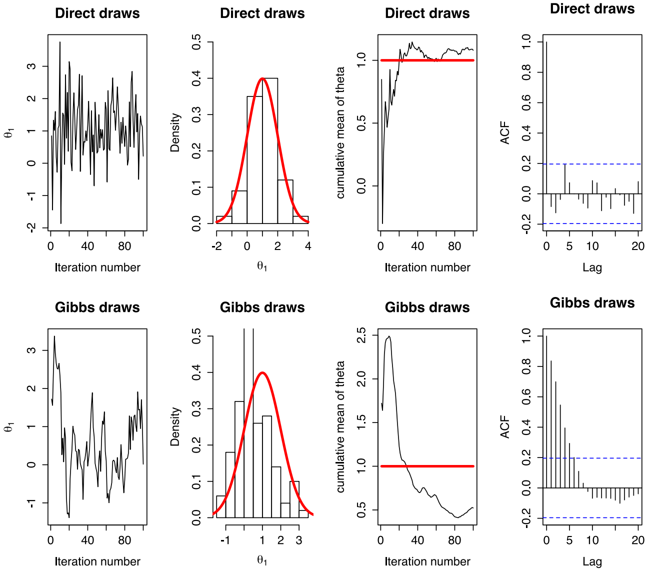


Gibbs sampling - Bivariate normal



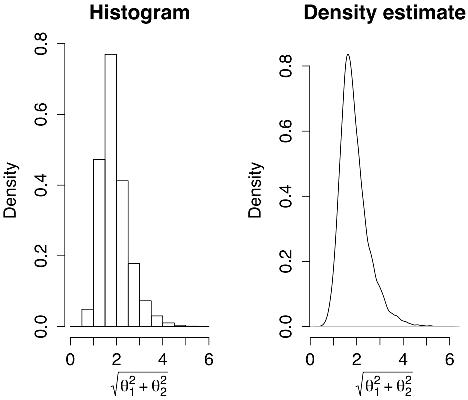
VOID

Direct sampling vs Gibbs sampling



Computing the density of functions of θ

- Given draws from the posterior $p(\theta_1, \theta_2 | Data)$ we can just compute the posterior of any function of θ_1 and θ_2. Example **impulse response functions**.

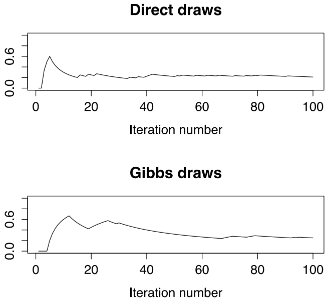


Computing complicated joint probabilities

- We can estimate a joint probability by counting:

$$Pr(\theta_1 > 0, \theta_2 > 0) \approx N^{-1} \sum_{i=1}^{N} 1(\theta_1^{(i)} > 0, \theta_2^{i)} > 0)$$

- We can for example easily compute $Pr(\text{inflation} > 3\% \textbf{ and } \text{repo} < 0 | \text{Data})$. At any horizons.



Gibbs sampling for AR processes

- **AR(p) process**

$$x_t = \mu + \phi_1(x_{t-1} - \mu) + ... + \phi_p(x_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2).$$

- Let $\phi = (\phi_1, ..., \phi_p)'$.
- **Prior**:
  - $\mu \sim$Normal
  - $\phi \sim$Multivariate Normal
  - $\sigma^2 \sim$Scaled Inverse χ^2.
- The **posterior** can be simulated by Gibbs sampling:
  - $\mu|\phi, \sigma^2, x \sim$ Normal
  - $\phi|\mu, \sigma^2, x \sim$ Multivariate Normal
  - $\sigma^2|\mu, \phi, x \sim$ Scaled Inverse χ^2

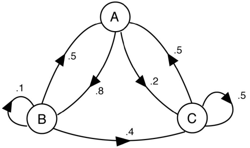
Markov chains

- Let $\mathcal{S} = \{s_1, s_2, ..., s_k)$ be a finite set of **states**.
  - Weather: $\mathcal{S} = \{\text{sunny}, \ \text{rain}\}$.
  - Journal rankings: $\mathcal{S} = \{A+, A, B, C, D, E\}$
- **Markov chain** is a stochastic process $\{X_t\}_{t=1}^{T}$ with random **state transitions**

$$p_{ij} = \Pr(X_{t+1} = s_j | X_t = s_i)$$

- Example realization journal ranking: $X_1 = C, X_2 = C, X_3 = B, X_4 = A+, X_5 = B$.
- **Transition matrix** for weather example

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix}$$



Stationary distribution

- *h*-step transition probabilities

$$P_{ij}^{(h)} = \Pr(X_{t+h} = s_j | X_t = s_i)$$

- *h*-step transition matrix

$$P^{(h)} = P^h$$

- The chain has a **unique equilibrium** stationary distribution $\pi = (\pi_1, ..., \pi_k)$ if it is
  - **irreducible** (possible to get from any state from any state)
  - **aperiodic** (does not get stuck in predictable cycles)
  - **positive recurrent** (expected time of returning to any state is finite)
- Limiting (long-run) distribution

$$P^t \rightarrow \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \rightarrow \infty$$

Stationary distribution, cont.

- Limiting (long-run) distribution

$$P^t \to \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \to \infty$$

- **Stationary distribution**

$$\pi = \pi P$$

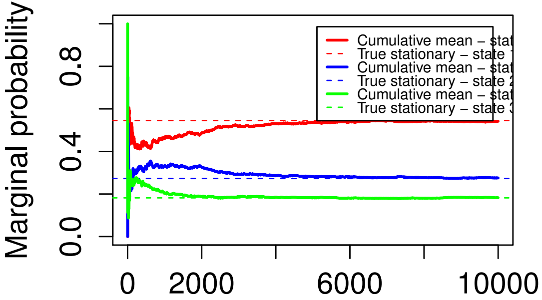
- Example:

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

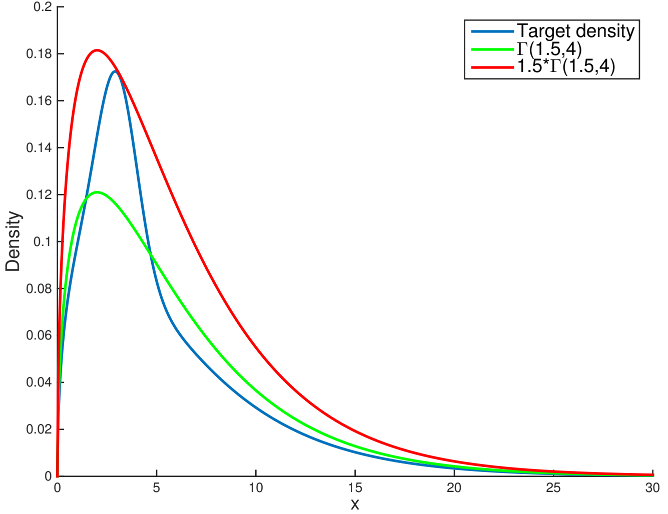
$$\pi = (0.545, 0.272, 0.181)$$

The basic MCMC idea

- Aim: to simulate from a discrete distribution $p(x)$ when $x \in \{s_1, s_2, ..., s_k\}$.
- **MCMC**: **simulate a Markov Chain** with a **stationary distribution** that is exactly $p(x)$.
- How to set up the transition matrix P? **Metropolis-Hastings**!



Rejection sampling



Random walk Metropolis algorithm

- Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, \ldots$

  1. Sample $\theta_p | \theta^{(i-1)} \sim N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$ (the **proposal distribution**)
  2. Compute the **acceptance probability**

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

  3. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Random walk Metropolis, cont.

- Assumption: we can compute $p(\theta_p|\mathbf{y})$ for any θ.
- Proportionality constant in $p(\theta_p|\mathbf{y})$ will cancel in α:

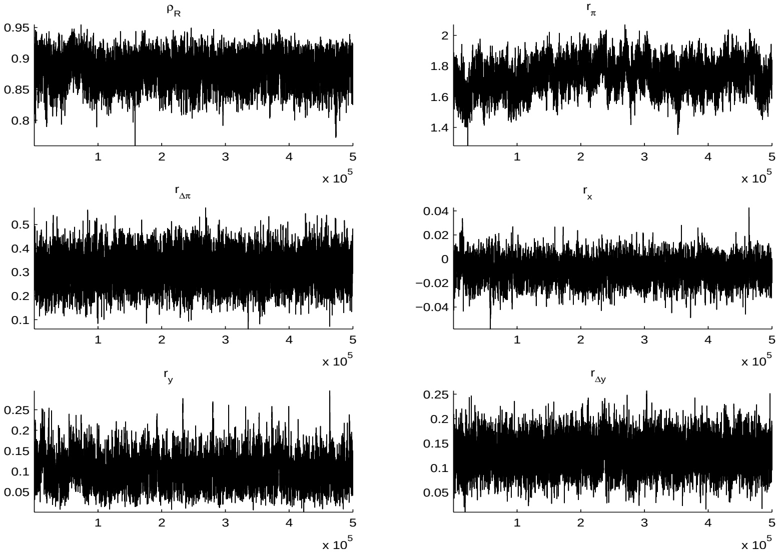
$$\alpha = \min\left(1, \frac{p\left(\mathbf{y}|\theta_p\right) p\left(\theta_p\right)}{p\left(\mathbf{y}|\theta^{(i-1)}\right) p\left(\theta^{(i-1)}\right)}\right)$$

- Just need to code up the loglikelihood $p\left(\mathbf{y}|\theta\right)$and the log prior $p\left(\theta\right)$.

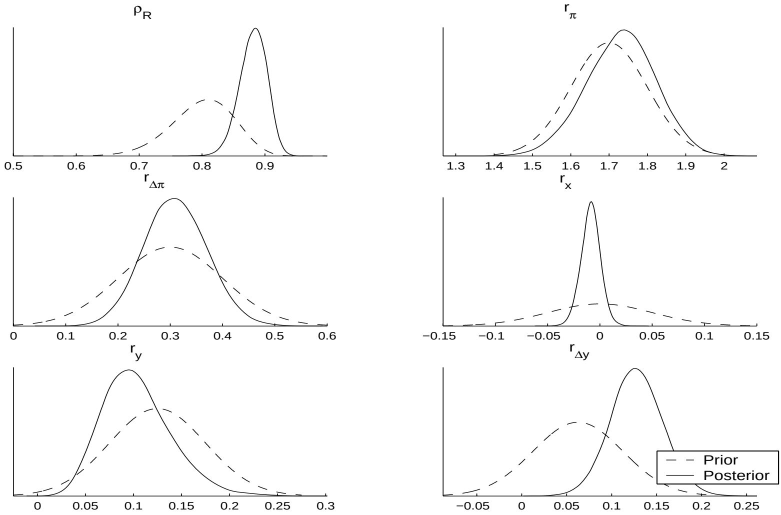
Random walk Metropolis, cont.

- Common choices of Σ in proposal $N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$:
  - $\Sigma = I$ (warning: may propose 'off the cigar'. Inefficient.)
  - $\Sigma = -H^{-1}$(Hessian of the log posterior evaluated at$\hat{\theta}_{mode}$) (propose 'along the cigar')
  - Adaptive. Start with $\Sigma = I$ and then recompute Σ from an initial simulation run.
- Get $\hat{\theta}_{mode}$ and H from numerical optimization (e.g. **fminunc** in Matlab).
- c is set so that average acceptance probability is roughly 25-35%.
- A good proposal:
  - should take reasonably **large steps** in θ-space
  - should **not be rejected too often**.

Random walk Metropolis for a DSGE [1]



Random walk Metropolis for a DSGE [1]



THE METROPOLIS-HASTINGS ALGORITHM

- Generalization when the proposal density is not symmetric.

- Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, \ldots$

1. Sample $\theta_p \sim q\left(\cdot|\theta^{(i-1)}\right)$ (the **proposal distribution**)

2. Compute the **acceptance probability**

$$\alpha = \min\left(1, \frac{p(\mathbf{y}|\theta_p)p(\theta_p)}{p(\mathbf{y}|\theta^{(i-1)})p(\theta^{(i-1)})}\frac{q\left(\theta^{(i-1)}|\theta_p\right)}{q\left(\theta_p|\theta^{(i-1)}\right)}\right)$$

3. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

The independence sampler

- **Independence sampler**: $q\left(\theta_p|\theta^{(i-1)}\right) = q\left(\theta_p\right)$.
- Proposal is independent of previous draw.
- Example:
$$\theta_p \sim t_v\left(\hat{\theta}, -H^{-1}\right),$$
where $\hat{\theta}$ and H^{-1} are computed by numerical optimization.
- Can be very **efficient**, but has a tendency to **get stuck**. **Parallelizable**!.
- Make sure that $q\left(\theta_p\right)$ has **heavier tails** than $p(\theta|\mathbf{y})$.

RWM and Indep Metropolis for a DSGE [2]

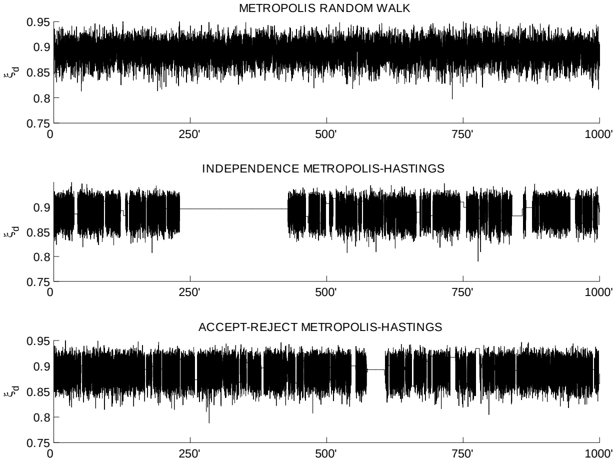


Figure 2. Sampling paths for the domestic sticky price parameter for three different posterior sampling algorithms.

The (in)efficiency of MCMC

- $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are **dependent** (autocorrelated).
- How efficient is my MCMC compared to iid sampling?
- If $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are iid with variance σ^2, then

$$\mathrm{Var}(\bar{\theta}) = \frac{\sigma^2}{N}.$$

- If $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are generated by MCMC

$$\mathrm{Var}(\bar{\theta}) = \frac{\sigma^2}{N}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$$

  where $\rho_k = Corr(\theta^{(i)}, \theta^{(i+k)})$ is the autocorrelation at lag k.
- **Inefficiency factor**

$$\mathrm{IF} = 1 + 2\sum_{k=1}^{\infty}\rho_k$$

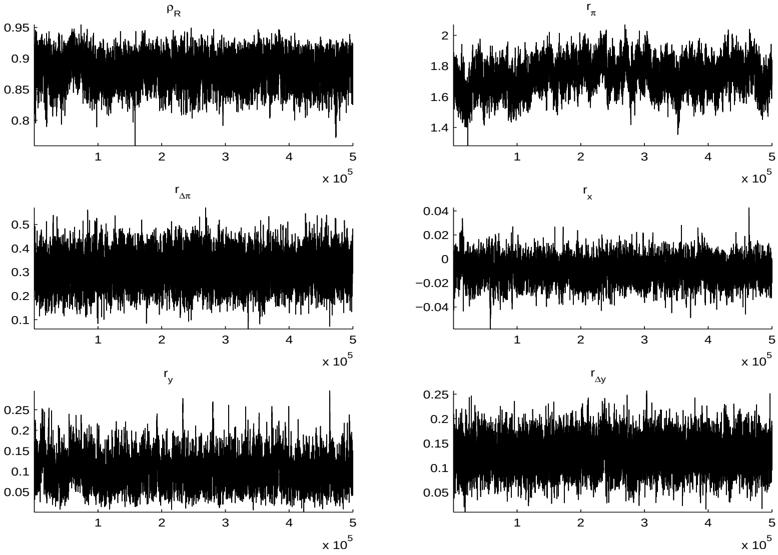
- **Effective sample size** from MCMC

$$\mathrm{ESS} = N/\mathrm{IF}$$

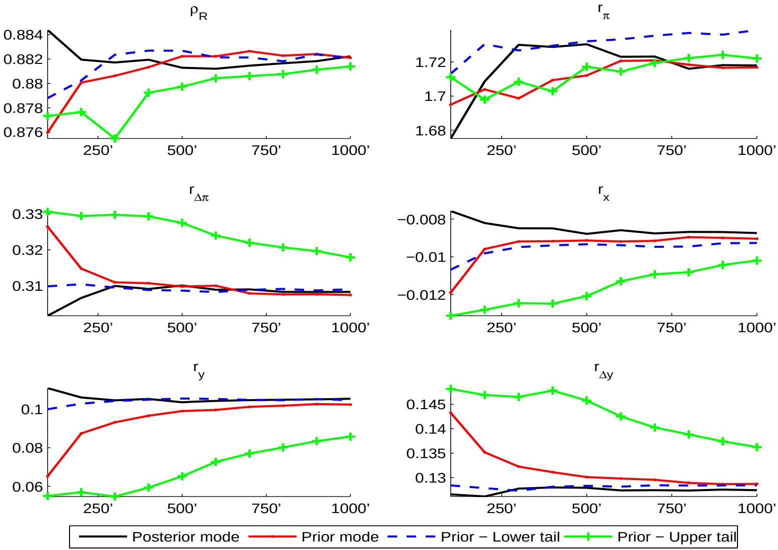
Burn-in and convergence

- **How long burn-in**?
- **How long to sample** after burn-in?
- To **thin** of not to thin? Only keeping every h draw reduces autocorrelation.
- **Convergence diagnostics**
  - Raw plots of simulated sequences (trajectories)
  - CUSUM plots + Local means
  - Potential scale reduction factor, R.

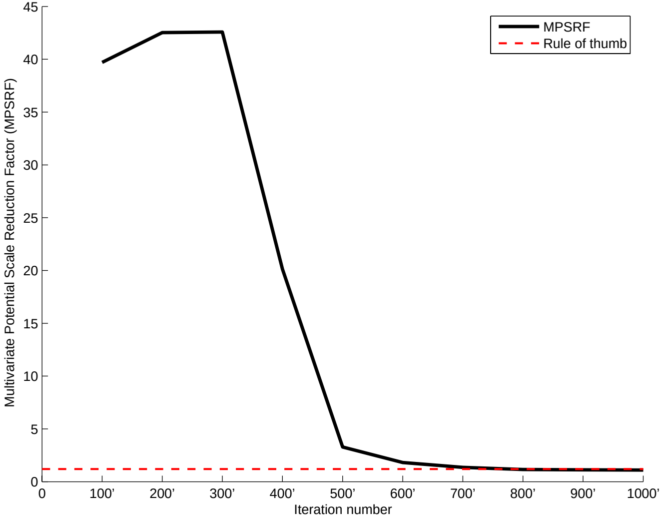
RANDOM WALK METROPOLIS FOR A DSGE [1]



RANDOM WALK METROPOLIS FOR A DSGE [1]



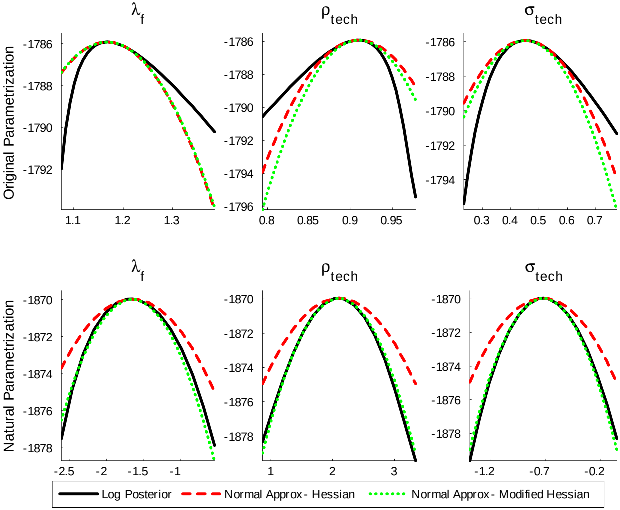
Potential Scale Reduction Factor - DSGE [1]



Transforming parameters

- Many models have parameters with **restricted parameter spaces**:
  - Stationary AR(1) process: $y_t = \rho y_{t-1} + \varepsilon_t$: $\rho \in [-1, 1]$ or $\rho \in [0, 1]$.
  - Variances: $\sigma^2 \in (0, \infty)$.
- **MCMC** and **optimization** is typically easier if we transform parameters to $(-\infty, \infty)$.
- Other advantages: often **improves normality** of the posterior. **Laplace approx** better.
- **Standard transformations**:
  - Parameters in $[0, 1]$: $\phi = \log\left(\frac{\rho}{1-\rho}\right)$
  - Parameters in $(0, \infty)$: $\phi = \log \sigma^2$.
- Don't forget the **Jacobian** of the transformation when computing the posterior density!

TRANSFORMING PARAMETERS - POSTERIOR NORMALITY



Transforming parameters - marginal likelihood

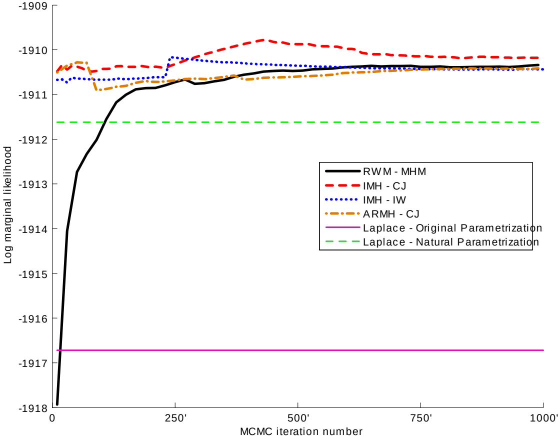


FIGURE 4. Sequential marginal likelihood estimates from the large-scale DSGE model for a subset of marginal likelihood estimators.

M. Adolfson, S. Laséen, J. Lindé, and M. Villani, "Bayesian estimation of an open economy dsge model with incomplete pass-through," *Journal of International Economics*, vol. 72, no. 2, pp. 481–511, 2007.

M. Adolfson, J. Lindé, and M. Villani, "Bayesian analysis of dsge models - some comments," *Econometric Reviews*, vol. 26, no. 2-4, pp. 173–185, 2007.