Chapter Seven

Markov Chain Monte Carlo Methods

Tingguo Zheng (zhengtg@gmail.com)
School of Economics & WISE & Chow, Xiamen University

(For Chow & SOE & WISE Msc and PhD Program, 2026-2027, Autumn Semester)

- **Reference:**
  - "Analysis of Financial Time Series", 2010, Tsay, Chapter 12.
  - Michael Johannes and Nicholas Polson, 2009, "Markov Chain Monte Carlo" in "Handbook of Financial Time Series", 2009, 1001-1012.

- **Outline:**
  - Basics of Bayesian inference
  - Classical sampling methods
  - MCMC Algorithms
  - MCMC for regression models
  - MCMC for univariate SV models
  - MCMC for TVP-VAR models

1 Basics of Bayesian Inference

In Bayes, the information brought by the data y, realization of $y \sim p(y|\theta)$, is combined with *prior information* specified by *prior distribution* with density $p(\theta)$.

Data information: In general, the *likelihood function* is defined as

$$L(\theta) = p(y|\theta).$$

Posterior distribution: The *posterior distribution* $\pi(\theta) := p(\theta|y)$ is derived from the joint distribution $p(y|\theta)p(\theta)$, according to

$$\pi(\theta) = p(\theta|y) = \frac{p(y|\theta)p(\theta)}{m(y)} \Leftrightarrow \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta},$$

where the *marginal density* of Y is given by

$$m(y) = \int p(y|\theta)p(\theta)d\theta.$$

Then, the *posterior distribution* of θ

$$\pi(\theta) \propto p(y|\theta)p(\theta) = \underbrace{L(\theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}}.$$

- Operates conditional upon the observations. Integrate simultaneously *prior information* and *data information*.

- One can take advantage of *subjective* (e.g. conjugate prior) and/or *nonsample* (e.g. $p(\theta) \propto 1$) prior information in Bayesian inference.

As may be expected, an *important goal* of the Bayesian analysis is to summarize the *posterior density*.

Particular summaries, such as the *posterior mean* and *posterior covariance matrix*, are especially important as are *interval estimates* (called credible intervals) with specified posterior probabilities.

1. The calculation of these quantities reduces to the evaluation of the following integral
$$\int h(\theta)\pi(\theta)d\theta,$$
under various choices of the function h.

2. For example, to get the *posterior mean*, one lets $h(\theta) = \theta$ and for the *second moment matrix* one lets $h(\theta) = \theta\theta'$, from which the *posterior covariance matrix* and *posterior standard deviations* may be computed.

Monte Carlo simulation

- Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ denote a random vector having a given density function $p(x_1, \ldots, x_n)$ and suppose we are interested in computing

$$E[h(\boldsymbol{X})] = \int \cdots \int h(x_1, \ldots, x_n) p(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

  for some n-dimensional function h.

- In many situations, it is not analytically possible either to compute the preceding multiple integral exactly or even to numerically approximate it within a given accuracy.

- One possibility that remains is to approximate $E[h(\boldsymbol{X})]$ by means of simulation.

- To approximate $E[h(\boldsymbol{X})]$, start by generating a random vector $\boldsymbol{X}^{(1)} = (X_1^{(1)}, \ldots, X_n^{(1)})$ from the corresponding joint density $p(x_1, \ldots, x_n)$ and then compute $Y^{(1)} = h(\boldsymbol{X}^{(1)})$.

- Then generate a second random vector (*independent* of the first) $\boldsymbol{X}^{(2)}$ and compute $Y^{(2)} = h(\boldsymbol{X}^{(2)})$.

- *Keep on doing* this until G, a fixed number, of *independent and identically distributed* random variables $Y^{(g)} = h(\boldsymbol{X}^{(g)})$, $i = 1, \ldots, G$ have been generated.

- Now by the *strong law of large numbers*, we know that

$$\lim_{G \to \infty} \frac{1}{G} \sum_{g=1}^{G} Y^{(g)} = E[Y] = E[h(\boldsymbol{X})].$$

MCMC methods

- To summarize the posterior density $p(\theta|y)$ one can produce a *simulated sample* $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$ from this posterior density.

- From this simulated sample, the *posterior expectation* of $h(\theta)$ can be estimated by the average

$$M^{-1} \sum_{j=1}^{M} h(\theta^j).$$

- In the context of MCMC sampling, a suitable *law of large numbers* for *Markov chains* that is presented below can be used establish the fact that

$$M^{-1} \sum_{j=1}^{M} h(\theta^j) \to \int h(\theta) p(\theta|y) d\theta, \quad M \to \infty.$$

2 Classical sampling methods

(1) Inverse transform method

This method is particularly useful in the context of *discrete distribution functions* and is based on taking the *inverse transform* of the *CDF*.

Generate discrete random variable

- Suppose the *mass function* of a discrete random variable

$$\Pr(x = x_j) = p_j, \quad j = 1, 2, \ldots, \quad \sum_j p_j = 1,$$

and *cumulative mass function*

$$\Pr(x \leq x_j) \equiv F(x_j) = p_1 + p_2 + \cdots + p_j.$$

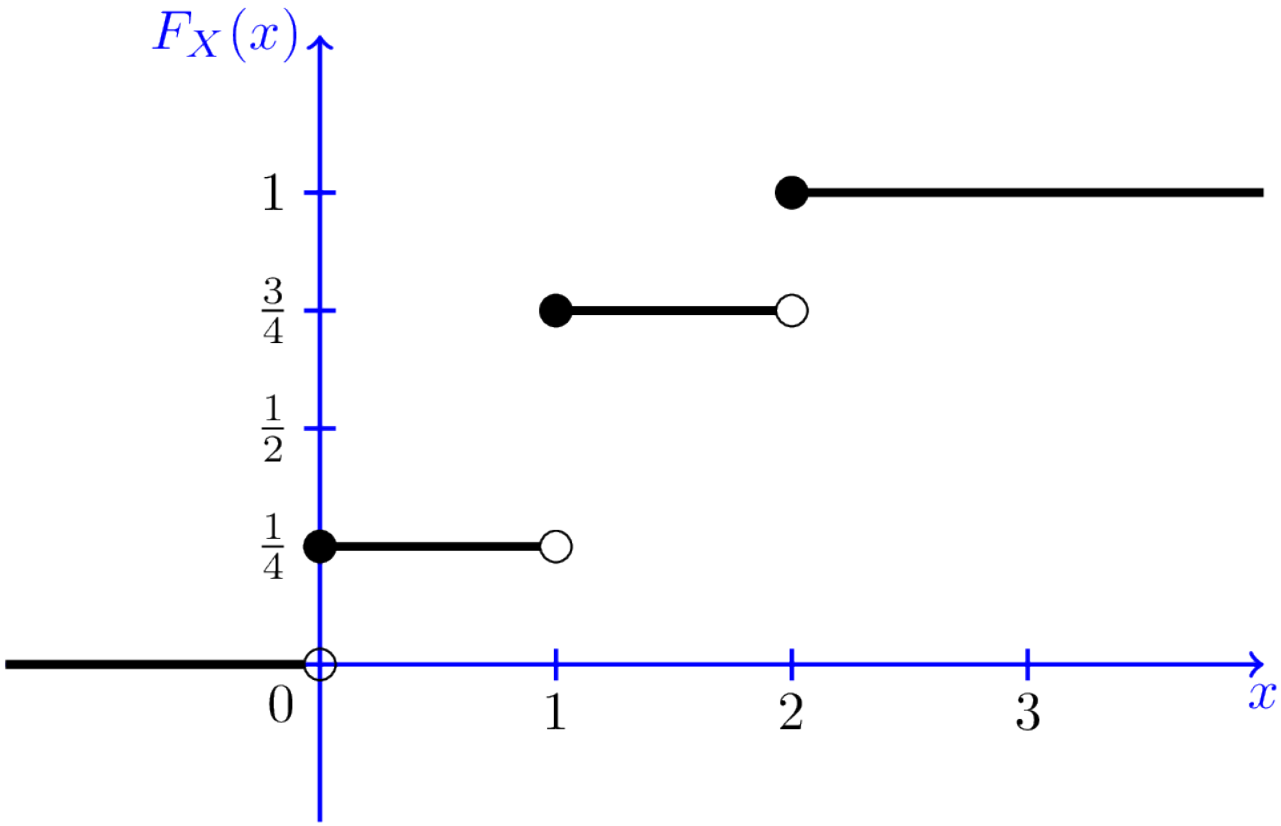


Figure 1 The Cumulative Distribution Function of a Discrete Random Variable.

- The function F is a right-continuous *step function* 阶跃函数 that has jumps at the point x_j equal to p_j and is constant otherwise.

- It is not difficult to see that its inverse takes the form

$$F^{-1}(u) = x_j \quad \text{if} \quad p_1 + \cdots + p_{j-1} < u \leq p_1 + \cdots + p_j. \tag{1}$$

- A random variate from this distribution is obtained by generating U *uniform* on $(0, 1)$ and computing $F^{-1}(U)$ where F^{-1} is the inverse function in Equation (1).

- This method samples x_i with probability p_j because

$$\begin{aligned} \Pr(F^{-1}(U) = x_j) &= F(x_j) - F(x_j - 0) \\ &= p_j. \end{aligned}$$

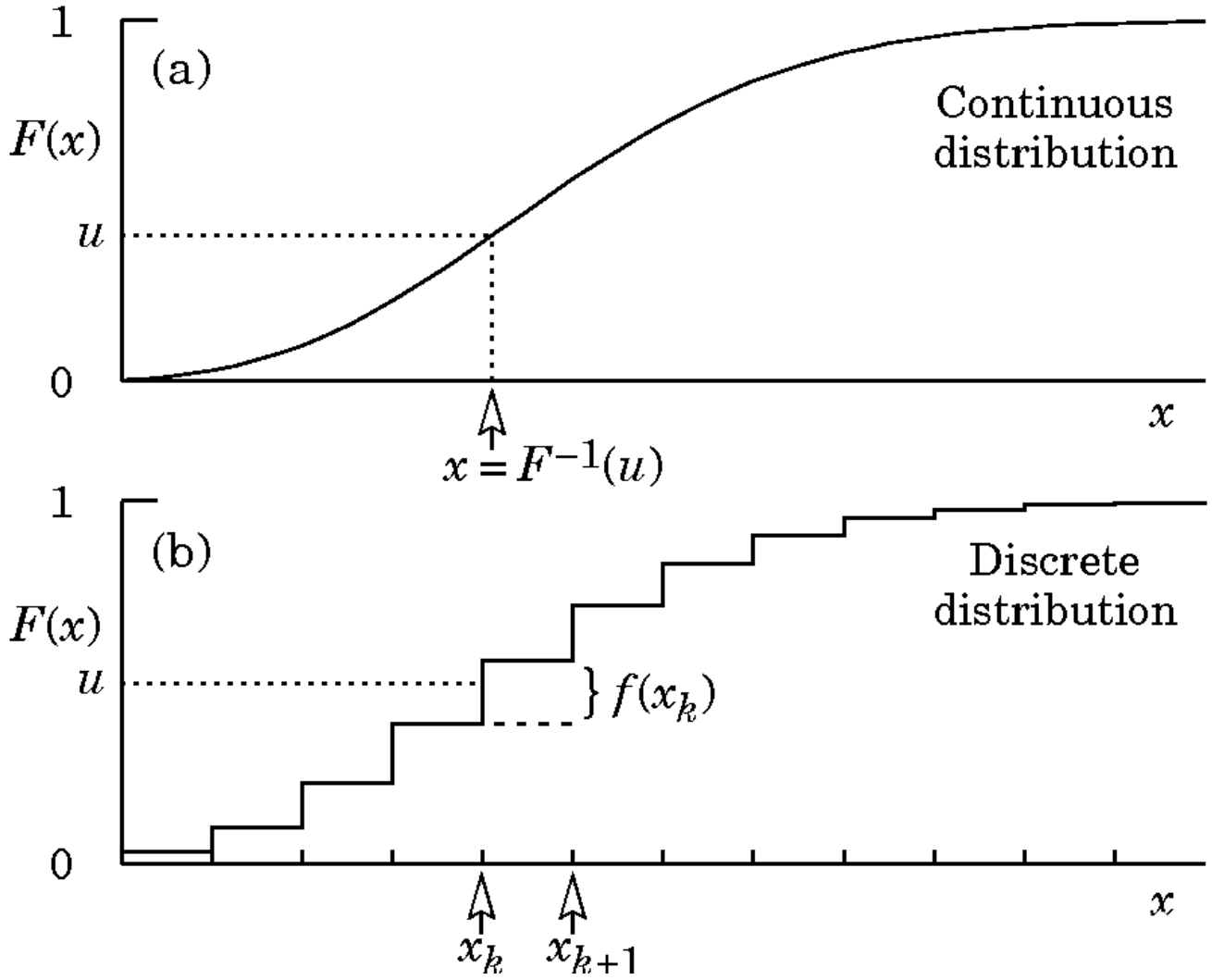


Figure 2 Generate Random Variable.

Continuous distribution random variable

- For a function F on $\mathbb{R}$, the *generalized inverse* of F, F^{-1}, is defined by

$$F^{-1}(u) = \inf\{x; F(x) \geq u\}.$$

- To generate a random variable $X \sim F$, simply

  1. first generate

$$U \sim \mathcal{U}_{[0,1]},$$

  2. then make the transform

$$x = F^{-1}(u).$$

Example 1 (Sampling an exponential random variable)**.**

- If $F(x) = 1 - e^{-x}$ for $x > 0$, then $F^{-1}(u)$ is that value of x such that

$$1 - e^{-x} = u \quad \Longrightarrow \quad x = -\log(1-u).$$

- Hence, if U is a *uniform* $(0, 1)$ variable, then

$$x = F^{-1}(U) = -\log(1-U)$$

  is *exponentially distributed* with *mean* 1.

- Since $1-U$ is also *uniformly distributed* on $(0, 1)$ it follows that $-\log U$ is *exponential* with *mean* 1.

(2) Accept-reject algorithm

- The *accept-reject* method is the basis for many of the well known *univariate random number generators*.

- This method is characterized by a *source density* (源密度) $q(x)$, which is used to supply *candidate values* (备选值), and a constant M that is determined by analysis, such that for all x

$$p(x) \leq Mq(x).$$

- One draws a variate from q, accepting it with probability $p(x)/Mq(x)$.

- If the particular *proposal* is rejected, a new one is drawn and the process continued until one is accepted.

Algorithm 1 (A/R algorithm).

1. Repeat for $i = 1, 2, \ldots, G$.

(a) Generate

$$x^{(i)} \sim q(x); \quad u \sim \mathcal{U}_{[0,1]}.$$

(b) If $u \leq \dfrac{p(x^{(i)})}{Mq(x^{(i)})}$, *then accept* $x^{(i)}$, *otherwise go to step 1(a).*

2. Return the values $\{x^{(1)}, x^{(2)}, \ldots, x^{(G)}\}$.

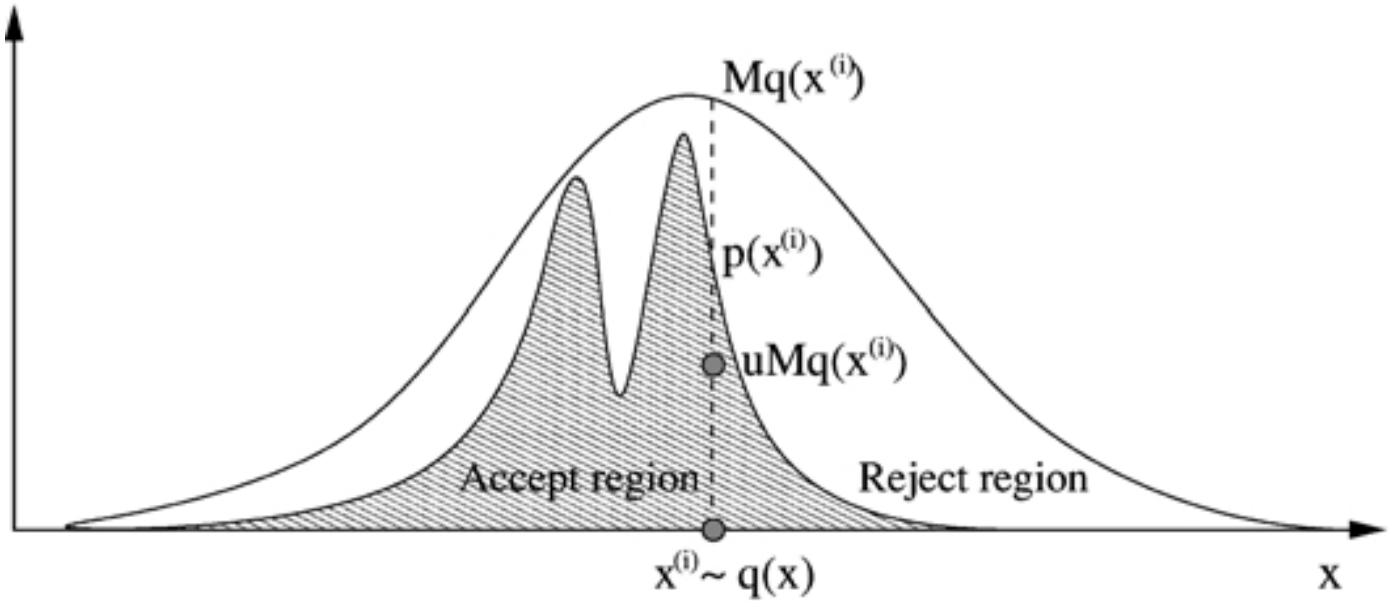


Figure 3 Rejection sampling: Sample a *candidate* $x^{(i)}$ and a *uniform variable* u. Accept the *candidate sample* if $uMq(x^{(i)}) < p(x^{(i)})$, otherwise reject it.

Example 2.

- Let us use the rejection method to generate a random variable having density function

$$f(x) = 20x(1-x)^3, \qquad 0 < x < 1.$$

- Since this random variable is concentrated in the interval $(0, 1)$, let us consider the rejection method with

$$g(x) = 1, \qquad 0 < x < 1.$$

- To determine the constant c such that $f(x)/g(x) \leq c$, we determine the *maximum value* of $\dfrac{d}{dx}\left[\dfrac{f(x)}{g(x)}\right] = 20[(1-x)^3 - 3x(1-x)^2]$. Setting

this equal to 0 shows that the *maximal value* is attained when $x = \frac{1}{4}$, and thus

$$\frac{f(x)}{g(x)} \leq 20 \left(\frac{1}{4}\right) \left(\frac{3}{4}\right)^3 = \frac{135}{64} \equiv c.$$

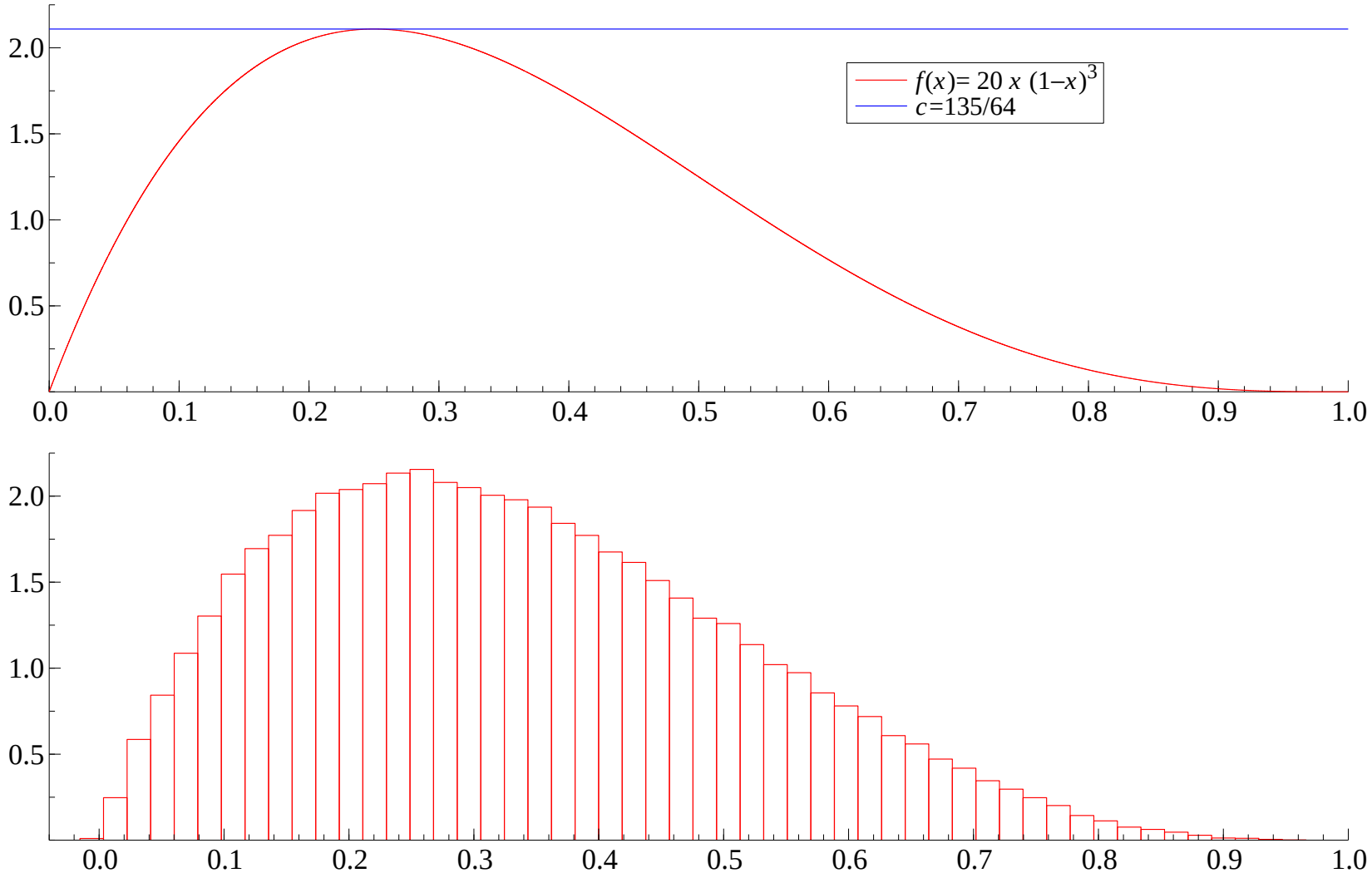
- Hence,

$$\frac{f(x)}{cg(x)} = \frac{256}{27} x(1-x)^3.$$

Thus the rejection procedure is as follows:

1. Generate *random numbers* $U_1 \sim \mathcal{U}_{[0,1]}$ (candidate) and $U_2 \sim \mathcal{U}_{[0,1]}$.
2. If $U_2 \leq \frac{256}{27} U_1 (1-U_1)^3$, stop and set $X = U_1$. Otherwise return to step 1.

See the next figure and my program.


$f(x)= 20\ x\ (1-x)^3$
c=135/64
2.0
1.5
1.0
0.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

3 MCMC Algorithms

3.1 Clifford-Hammersley theorem

- The *Clifford-Hammersley theorem* (CH) proves that the joint distribution, $p(\theta_1, \theta_2)$, is completely determined by the conditional distributions, $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$, under a positivity condition.

- CH implies that the same information can be extracted from the *lower-dimensional* conditional distributions, breaking *"curse of dimensionality"* by transforming a *higher dimensional* problem, sampling from $p(\theta_1, \theta_2)$, into easier problems, sampling from $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$.

- The general version of CH follows by analogy. Partitioning a vector as $\theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_K)$, then the general CH theorem states that

$$p(\theta_i|\theta_{-i}) \triangleq p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K),$$

for $i = 1, \ldots, K$, completely characterizes the joint distribution $p(\theta_1, \ldots, \theta_K)$.

Models with latent variables

- An important case arises frequently in models with *latent variables*.
- Here, the posterior is defined over vectors of *static fixed parameters*, θ, and *latent variables*, x. In this case, CH implies that $p(\theta, x|y)$ is completely characterized by $p(\theta|x, y)$ and $p(x|\theta, y)$.
- The distribution $p(\theta|x, y)$ is the *posterior distribution* of the parameters, conditional on the observed data and the latent variables.
- Similarly, $p(x|\theta, y)$ is the *smoothing distribution* of the latent variables given the parameters.

3.2 Metropolis Hastings

- The *Metropolis-Hastings* (MH) algorithm provides a general approach for sampling from a given *target density*, $\pi(\theta) := p(\theta|y)$.

- MH uses an *accept-reject approach*, drawing a *candidate* (备选) from a distribution q that is accepted or rejected based on an *acceptance probability*.

- *Unlike traditional accept-reject algorithms, which repeatedly sample until acceptance, the MH algorithm samples only once at each iteration.* If the *candidate* is rejected, the algorithm keeps the current value. So the entire vector θ is updated at once.

Algorithm 2 (Metropolis-Hastings)**.**

1. Specify an initial value $\theta^{(0)}$*;*

2. Repeat for $g = 1, \ldots, G$*.*

(a) Draw a value θ' *from a proposal distribution,* $q(\theta'|\theta^{(g)})$

(b) Take

$$\theta^{(g+1)} = \begin{cases} \theta', & \text{with probability } \alpha(\theta^{(g)}, \theta'); \\ \theta^{(g)}, & \text{with probability } 1 - \alpha(\theta^{(g)}, \theta'). \end{cases}$$

where

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

3. Return the values $\{\theta^{(1)}, \ldots, \theta^{(G)}\}$*.*

- To implement the *accept-reject* step, draw a *uniform random variable*, $U \sim \mathcal{U}_{[0,1]}$, and set $\theta^{(g+1)} = \theta'$ if $U < \alpha(\theta^{(g)}, \theta')$, otherwise leaving $\theta^{(g)}$ unchanged ($\theta^{(g+1)} = \theta^{(g)}$).

- It is important to note that the denominator in the *acceptance probability* cannot be zero, provided the algorithm is started from a π-positive point since q is always positive.

- The MH algorithm only requires that π can be evaluated up to proportionality.

- The output of the algorithm, $\{\theta^{(g)}\}_{g=1}^{G}$, is clearly a *Markov chain*. The key theoretical property is that the Markov chain, under mild regularity, has $\pi(\theta)$ as its *limiting distribution*.

Transition kernel

Through a *proposal distribution* q, the MH algorithm defines a *Markov Chain* with the following *transition kernel*

$$K(x, y) = \alpha(x, y)q(y|x) + \delta_x(y)r(x),$$

where $r(x) = \int (1 - \alpha(x, y))q(y|x)dy = 1 - \int \alpha(x, y)q(y|x)dy$ and $\delta_x(y)$ is the *Dirac measure* corresponding to a *indicator function* ($\delta_x(y) = 1$ if $x = y$, $\delta_x(y) = 0$ otherwise).

The transition kernel is composed of two terms. The *first term* represents *the probability of arriving in some* $y \neq x$. The *second term* corresponds to *the probability of remaining in state* x, i.e. $y = x$, where $r(x)$ encodes *the probability of rejecting a move*.

Time reversibility: The goal is to check if $K(x,y)\pi(x) = K(y,x)\pi(y)$.

For the first term, the *detailed balance condition* holds because

$$\begin{aligned}
\alpha(x,y)q(y|x)\pi(x) &= \min\left\{\frac{\pi(y)}{\pi(x)}\frac{q(x|y)}{q(y|x)}, 1\right\} q(y|x)\pi(x) \\
&= \min\left\{\pi(y)q(x|y), q(y|x)\pi(x)\right\} \\
&= \min\left\{1, \frac{\pi(x)}{\pi(y)}\frac{q(y|x)}{q(x|y)}\right\} q(x|y)\pi(y) \\
&= \alpha(y,x)q(x|y)\pi(y),
\end{aligned}$$

while the second term of the *transition kernel* satisfies that

$$r(x)\delta_x(y)\pi(x) = r(y)\delta_y(x)\pi(y)$$

since due to the multiplication with the Dirac measure, both sides of the equation are not zero only when $x = y$.

Convergence

There are two subtle forms of *convergence* operating:

- the convergence of the sample average.
- the distributional convergence of the chain.

Theorem 1 (**Ergodic Averaging).** *Suppose* $\theta^{(g)}$ *is an ergodic chain with stationary distribution* π *and suppose* f *is a real-valued function with* $\int |f| d\pi < \infty$*. Then for all starting points* θ

$$\lim_{G\to\infty} \frac{1}{G} \sum_{g=1}^{G} f(\theta^{(g)}) = \int f(\theta)\pi(\theta)d\theta = \mathrm{E}[f(\theta)]$$

almost surely.

In many cases, we go further with an ergodic central limit theorem:

Theorem 2 (**Central Limit Theorem).** *Suppose* $\theta^{(g)}$ *is an ergodic chain with stationary distribution* π *and suppose that* f *is real-valued and* $\int |f| d\pi < \infty$. *Then there exists a real number* $\sigma(f)$ *such that*

$$\sqrt{G}\left(\frac{1}{G}\sum_{g=1}^{G} f(\theta^{(g)}) - \int f(\theta)\pi(\theta)d\theta\right) \quad \Rightarrow N(0, \sigma^2(f))$$

converges in distribution to a mean zero normal distribution with variance $\sigma^2(f)$ *for any starting value.*

(1) Independence MH——Hastings (1970)

One special case draws a candidate independently of the previous state, $q(\theta'|\theta^{(g)}) = q(\theta')$. In this *independence* MH algorithm, the *acceptance criterion* simplifies to

$$\alpha\left(\theta^{(g)}, \theta'\right) = \min\left(\frac{\pi(\theta')}{\pi(\theta^{(g)})}\frac{q(\theta^{(g)})}{q(\theta')}, 1\right).$$

- When using independence MH, it is common to pick the *proposal density* to closely match certain properties of the target distribution.
  - One common criterion is to ensure that tails of the *proposal density* are thicker than the tails of the *target density*.
  - By *"blanketing"* the target density, it is less likely that the Markov chain will get trapped in a low probability region of the state space.

(2) Random-walk Metropolis——Metropolis et al. (1953)

- *Random-walk* (RW) Metropolis is the polar opposite of the independence MH algorithm. It draws a candidate from the following RW model,

$$\theta' = \theta^{(g)} + \sigma\varepsilon_t,$$

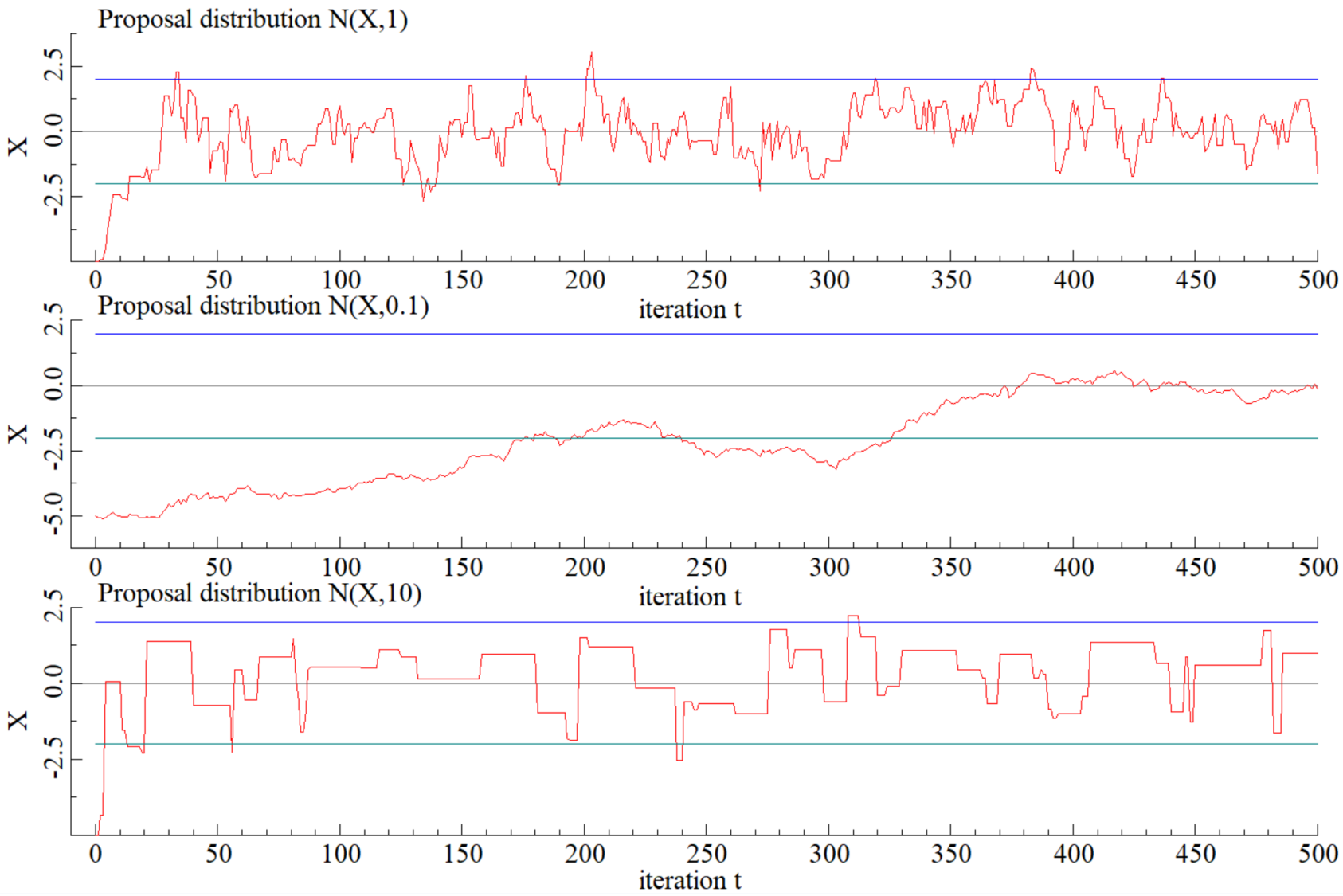
where ε_t is an *independent*, mean zero, and *symmetric* error term, typically taken to be a normal or t-distribution, and σ is a *scaling factor*.

- *Symmetry* ($q(z_t) = q(-z_t)$, where $z_t = \sigma\varepsilon_t$) implies that $q(\theta'|\theta^{(g)}) = q(\theta^{(g)}|\theta')$, with *acceptance probability*

$$\alpha\left(\theta^{(g)}, \theta'\right) = \min\left(\pi(\theta')/\pi(\theta^{(g)}), 1\right).$$

- The algorithm must be *tuned*, by adjusting the variance of the error term, to obtain an *acceptable level* of accepted draws, generally in *the range* of $20 - 40\%$.
  - If it is too large it is possible that the chain may remain stuck at a particular value for many iterations.
  - If it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution.
  - Both circumstances will tend to generate draws that are highly serially correlated.
  - See next figure.

- The RW algorithm, unlike the independence algorithm, learns about the density $\pi(\theta)$ via small *symmetric steps*, randomly "walks" around the support of π.


Proposal distribution N(X,1)
X
2.5
0.0
-2.5
0
50
100
150
200
250
300
350
400
450
500
iteration t
Proposal distribution N(X,0.1)
X
2.5
0.0
-2.5
-5.0
0
50
100
150
200
250
300
350
400
450
500
iteration t
Proposal distribution N(X,10)
X
2.5
0.0
-2.5
0
50
100
150
200
250
300
350
400
450
500
iteration t

(3) Tailed MH——Chib and Greenberg (1994)

- They discuss a way of formulating *proposal densities* in the context of time series *autoregressive-moving average models* that has a bearing on the choice of proposal density for the independence M-H chain.

- They suggest matching the proposal density to the target at the mode by a *multivariate normal* or *multivariate-t* distribution with location given by the mode of the target and the dispersion given by *inverse of the Hessian* evaluated at the mode.

- Specifically, the parameters of the *proposal density* are taken to be

$$\boldsymbol{m} = \operatorname{argmax} \log \pi(\theta) \quad \text{and} \quad \boldsymbol{V} = \tau \left\{ -\frac{\partial^2 \log \pi(\theta)}{\partial \theta \partial \theta'} \right\}^{-1}_{\theta = \widehat{\theta}},$$

where τ is a *tuning parameter* that is adjusted to control the *acceptance rate*.

- The proposal density is then specified as $q(\theta') = f(\theta'|\boldsymbol{m}, \boldsymbol{V})$, where f is some *multivariate density*.

3.3 Multiple-block MH algorithm——Hastings (1970)

- When the dimension of θ is *quite large* it is preferable to construct the Markov chain simulation by first grouping the variables θ into p *blocks* $(\theta_1, \ldots, \theta_p)$, and sampling each block, conditioned on the rest, by the M-H algorithm.

- Let $\theta_{-k} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_p)$ denote the variables (blocks) excluding θ_k.

- Also let $\pi(\theta_k, \theta_{-k})$ denote the joint density of θ, regardless of where θ_k appears in the list $(\theta_1, \ldots, \theta_p)$.

- Furthermore, let $\{q_k(\theta_k'|\theta_k, \theta_{-k})\}_{k=1}^p$ denote a *collection* of proposal densities, *one for each blocks*.

- Finally, define

$$\alpha_k(\theta_k, \theta_k'|\theta_{-k}) = \min\left\{\frac{\overbrace{\pi(\theta_k', \theta_{-k})}^{\pi(\theta_k'|\theta_{-k})}}{\underbrace{\pi(\theta_k, \theta_{-k})}_{\pi(\theta_k|\theta_{-k})}} \frac{q_k(\theta_k|\theta_k', \theta_{-k})}{q_k(\theta_k'|\theta_k, \theta_{-k})}, 1\right\}$$

  as the probability of move for block θ_k conditioned on θ_{-k}.

- Then, in the *multiple-block M-H algorithm*, one cycle of the algorithm is completed by *updating each block*, say *sequentially in fixed order*, using a M-H step with the above probability of move, given the most current value of the remaining blocks.

- The algorithm may be summarized as follows.

Algorithm 3 (Multiple-block Metropolis-Hastings).

1. Specify an initial value $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.

2. Repeat for $g = 1, 2, \ldots, G$, *and* $k = 1, 2, \ldots, p$

(a) Draw $\theta_k' \sim q_k(\theta_k'|\theta_k^{(g)}, \theta_{-k})$

(b) Calculate

$$\alpha_k(\theta_k^{(g)}, \theta_k'|\theta_{-k}) = \min\left\{\frac{\pi(\theta_k', \theta_{-k})}{\pi(\theta_k^{(g)}, \theta_{-k})}\frac{q(\theta_k^{(g)}|\theta_k', \theta_{-k})}{q(\theta_k'|\theta_k^{(g)}, \theta_{-k})}, 1\right\}$$

(c) Set

$$\theta_k^{(g+1)} = \begin{cases} \theta_k', & \text{if } \mathcal{U}_{[0,1]} \leq \alpha_k(\theta_k^{(g)}, \theta_k'|\theta_{-k}); \\ \theta_k^{(g)}, & \text{otherwise.} \end{cases}$$

3. Return the values $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}\}$.

- First, the version of the algorithm presented above assumes that the blocks are *revised sequentially in fixed order rather than in random order*.

- Second, at the moment block k is updated in this algorithm, the blocks $(\theta_1, \ldots, \theta_{k-1})$ have *already been revised* while the blocks $(\theta_{k+1}, \ldots, \theta_p)$ have *not*. Thus, at each step of the algorithm one must be sure to *condition on the most current value of the blocks* in θ_{-k}.

- Finally, if the *proposal density* q_k is determined by tailoring to $\pi(\theta_k, \theta_{-k})$, as in Chib and Greenberg (1994), then this implies that *the proposal density is not fixed but varies across iterations*.

3.4 **Gibbs sampling**——Gelfand and Smith (1990)

- ★ The *Gibbs sampler*, which is a special case of the *multiple-block Metropolis-Hastings* method, simulates multidimensional posterior distributions by iteratively sampling from the lower-dimensional *conditional posteriors*.

- ★ The parameters are grouped into p blocks $(\theta_1, \ldots, \theta_p)$ and *each block* is sampled according to the *full conditional distribution* of block θ_k, defined as the conditional distribution under π of θ_k given all the other blocks θ_{-k} and denoted as $\pi(\theta_k|\theta_{-k})$.

- Similar to the *multiple-block M-H* algorithm, the most current value of the remaining blocks *is used in deriving* the full conditional distribution of each block.

- Derivation of the full conditional distributions is usually quite simple since, by *Bayes theorem*, $\pi(\theta_k|\theta_{-k}) \propto \pi(\theta_k, \theta_{-k})$, the joint distribution of *all the blocks*.

- ★ To define the *Gibbs sampling* algorithm, let the set of full conditional distributions be

$$\{\pi(\theta_1|\theta_2, \ldots, \theta_p), \pi(\theta_2|\theta_1, \ldots, \theta_p), \ldots, \pi(\theta_p|\theta_1, \ldots, \theta_{p-1})\}.$$

  - Now one cycle of the Gibbs sampling algorithm is completed by *simulating* $\{\theta_k\}_{k=1}^p$ *from these distributions*, recursively updating the conditioning variables as one moves through each distribution.
  - When $p = 2$ one obtains the *two block Gibbs sampler* that is featured in the work of Tanner and Wong (1987).

Algorithm 4 (Gibbs sampling).

1. Specify an initial value $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.

2. Repeat for $g = 1, 2, \ldots, G$,

(a) Draw $\theta_1^{(g+1)} \sim \pi(\theta_1 | \theta_2^{(g)}, \theta_3^{(g)}, \ldots, \theta_p^{(g)})$.
(b) Draw $\theta_2^{(g+1)} \sim \pi(\theta_2 | \theta_1^{(g+1)}, \theta_3^{(g)}, \ldots, \theta_p^{(g)})$.
(c) ⋮
(d) Draw $\theta_p^{(g+1)} \sim \pi(\theta_p | \theta_1^{(g+1)}, \theta_2^{(g+1)}, \ldots, \theta_{p-1}^{(g+1)})$.

3. Return the values $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}\}$.

Thus, the transition of θ_k from $\theta_k^{(g)}$ to $\theta_k^{(g+1)}$ is effected by taking a draw from the conditional distribution $\pi(\theta_k | \theta_1^{(g+1)}, \ldots, \theta_{k-1}^{(g+1)}, \theta_{k+1}^{(g)}, \ldots, \theta_p^{(g)})$.

Connection with the multiple-block M-H algorithm

- Now recall that the probability of move in the *multiple-block M-H algorithm* is

$$\alpha_k(\theta_k, \theta_k'|\theta_{-k}) = \min\left\{\frac{\pi(\theta_k', \theta_{-k})}{\pi(\theta_k, \theta_{-k})}\frac{q_k(\theta_k|\theta_k', \theta_{-k})}{q_k(\theta_k'|\theta_k, \theta_{-k})}, 1\right\}$$

- so if one substitutes

$$q_k(\theta_k|\theta_k', \theta_{-k}) = \pi(\theta_k, \theta_{-k}) \quad \text{and} \quad q_k(\theta_k'|\theta_k, \theta_{-k}) = \pi(\theta_k', \theta_{-k})$$

  in this expression all the terms cancel implying that the probability of accepting the proposal is one, i.e. $\alpha_k(\theta_k, \theta_k'|\theta_{-k}) = 1$.

- Thus, the Gibbs sampling algorithm is a special case of the *multipleblock M-H algorithm*.

3.5 Griddy Gibbs Sampler

- The *Griddy Gibbs sampler* is an approximation that can be applied to approximate the conditional distribution by a discrete set of points.

- Suppose that θ is continuously distributed and univariate and that $\pi(\theta)$ can be evaluated on a point by point basis, but that the distribution $\pi(\theta)$ is nonstandard and direct draws are not possible.

- The Griddy Gibbs sampler approximates the continuously distributed θ with a discrete mass of $N-$points, $\{\theta_j\}_{j=1}^{N}$.

- Given this approximation, Ritter and Tanner (1991) suggest the following algorithm:

1. Compute $\pi(\theta_j)$ for $j = 1, \ldots, N$ and set $w_j = \pi(\theta_j)$;
2. Normalize the weights to add to unity and use these weights to approximate the inverse CDF of $\pi(\theta)$;
3. Generate a sample from the approximate distribution by drawing a uniform on $[0, 1]$ and inverting the CDF.

- Ritter and Tanner (1991) discuss issues involved with the choice of grid of points (N) and show that this algorithm can provide accurate characterization of the conditional distribution in certain cases.

- In general, the algorithm performs well when the discretization is performed on a small number of parameters. In high dimensional systems, the algorithm is not likely to perform extremely well.

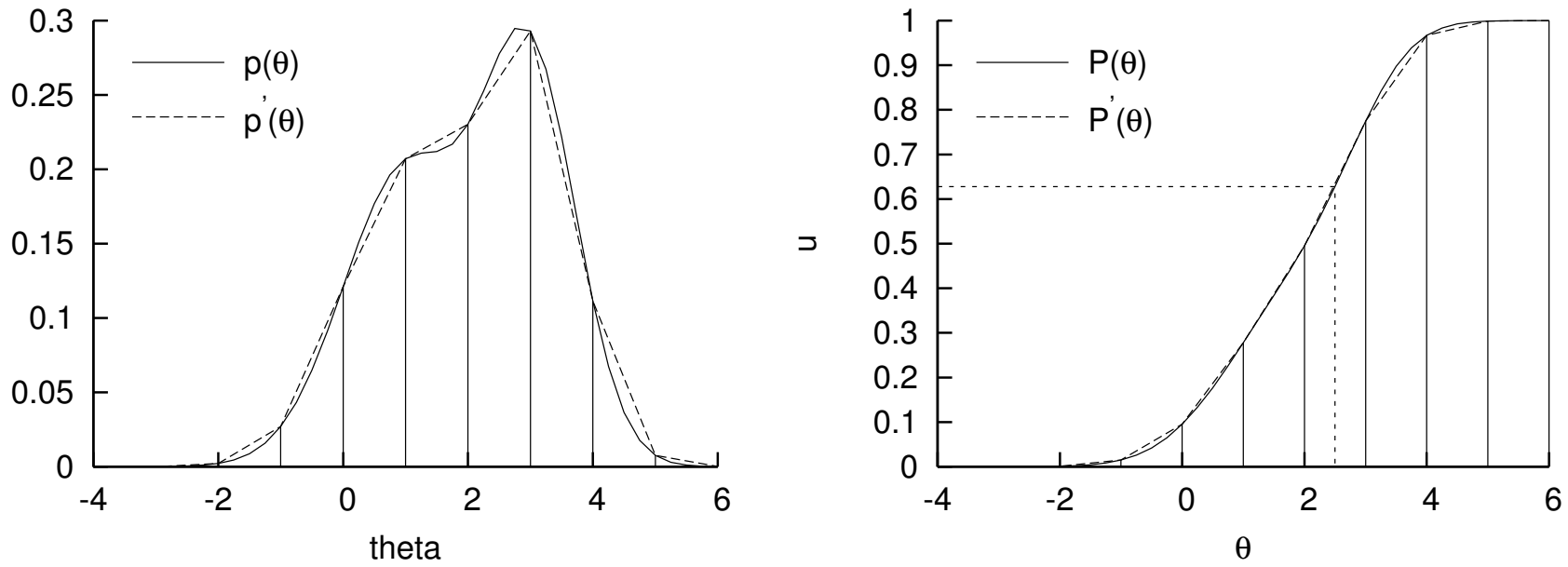


Figure 4 Sampling using the Griddy Gibbs sampler

Note: $p(\theta) = \pi(\theta)$, and $p'(\theta)$ is an approximation of $p(\theta)$.

3.6 Hybrid (混合) MCMC algorithm

- Given a partition of the vector θ via CH, a *hybrid MCMC algorithm* updates the chain one subset at a time, either by direct draws (*Gibbs steps*) or via MH (*Metropolis step*).

- Thus, a hybrid algorithm combines the features of the *MH algorithm* and the *Gibbs sampler*, providing significant flexibility in designing MCMC algorithms for different models.

To see the mechanics, consider the two-dimensional example.

- First, assume that the distribution $\pi(\theta_2|\theta_1)$ is recognizable and can be directly sampled with *Gibbs sampler*.

- Second, suppose that $\pi(\theta_1|\theta_2)$ can only be evaluated and not directly sampled. Thus we use a *Metropolis step* to update θ_1 given θ_2.
  - For the MH step, the *candidate* is drawn from $q(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)})$, which indicates that the step can depend on the *past draw* for θ_1.
  - We can draw $\theta_1^{(g+1)} \sim q\left(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)}\right)$ and then *accept/reject* based on

$$\alpha\left(\theta_1^{(g)}, \theta_1'\right) = \min\left(\frac{\pi(\theta_1'|\theta_2^{(g)})}{\pi(\theta_1|\theta_2^{(g)})}\frac{q(\theta_1|\theta_1', \theta_2^{(g)})}{q(\theta'|\theta_1^{(g)}, \theta_2^{(g)})}, 1\right).$$

Algorithm 5 (General hybrid algorithm).

1. Specify an initial value $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$.

2. Repeat for $g = 1, 2, \ldots, G$,

(a) Draw $\theta_1^{(g+1)} \sim q\left(\theta_1 | \theta_1^{(g)}, \theta_2^{(g)}\right)$ *by MH.*

(b) Draw $\theta_2^{(g+1)} \sim p(\theta_2 | \theta_1^{(g+1)})$ *by Gibbs sampler.*

3. Return the values $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}\}$.

- In *higher dimensional* cases, a *hybrid algorithm* consists of any combination of Gibbs and Metropolis steps.
- Hybrid algorithms significantly *increase* the applicability of MCMC methods.

3.7 On convergence of Chain*

For assessing convergence, several statistics have been devised (Geweke 1992, Kim, Shephard and Chib 1998, Yu and Mykland 1998).

(1) CUSUM plot

Yu and Mykland (1998) proposes to consider a plot of the CUSUM path with

$$\mathrm{CUSUM}_t = \frac{1}{t\sigma_\theta}\sum_{i=1}^{t}(\theta^{(i)} - \mu_\theta), \qquad t = 1, \ldots, N,$$

where μ_θ and σ_θ are the empirical mean and standard deviation of the complete sample.

A plot of CUSUM_t against time which diverges from zero for a prolonged period of time is an indication of bad convergence.

(2) Relative Numerical Efficiency (RNE) - Geweke1992

Geweke (1992) promotes the use of the Relative Numerical Efficiency (RNE) as a measure for the quality of a correlated sample.

It compares the empirical variance of the sample with a correlation-consistent variance estimator,

$$\mathrm{RNE} = \frac{\sigma_\theta^2}{\sigma_{\mathrm{NW},t}^2},$$

where σ_θ^2 is a direct estimator of the variance, and $\sigma_{\mathrm{NW},t}^2$ is the Newey-West (Newey and West 1987) variance estimator taking the correlation up to lags of $q\%$ of the size of the sample into account.

Practical values for q can be 4, 8 or even 15%.

(3) Relative Numerical Efficiency (RNE) - KSC1998

The implementation of Kim et al. (1998) computes the RNE as

$$\widehat{R}_{B_m} = 1 + \frac{2B_m}{B_m - 1} \sum_{i=1}^{B_m} K(i/B_m)\hat{\rho}(i),$$

with $K(j)$ the Parzen kernel, B_m the bandwidth and $\hat{\rho}(i)$ the i-th order autocorrelation.

This implementation can found in an Ox package named ”MCMCPack”.

(4) Mahalanobis distance

Another practical method of assessing convergence is to monitor the estimates of location and scale.

Assume an initial estimate of the location μ_0 is given, and that the sample run leads to estimates $\hat{\mu}$, $\widehat{\Sigma}$ of location and scale.

The Mahalanobis distance between the previous and present location is

$$d_{\text{Math}} = (\hat{\mu} - \mu_0)\widehat{\Sigma}^{-1}(\hat{\mu} - \mu_0)$$

When d_{Math} is close to zero, the algorithm has probably found the stable density.

4 MCMC for regression models

4.1 Univariate regression models

- Consider a *univariate* regression model:

$$Y = X\beta + \varepsilon$$

where Y is a $T \times 1$ observational vector, β is a $K \times 1$ vector and X is a $T \times K$ matrix of observations and we assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.

- Assuming standard *conjugate independent priors* for the parameters

$$p(\beta) = \mathcal{N}(\beta_0, \Sigma_0) = (2\pi)^{-K/2}|\Sigma_0|^{-1/2} \exp\left\{-\frac{1}{2}(\beta - \beta_0)'\Sigma_0^{-1}(\beta - \beta_0)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(\beta - \beta_0)'\Sigma_0^{-1}(\beta - \beta_0)\right\}$$

and the variance [1]

$$p(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}+1} \exp\left(-\frac{\delta_0}{2\sigma^2}\right).$$

- The *likelihood function* is given by:

$$p(Y|X, \beta, \sigma^2) \propto (\sigma^2)^{-\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right).$$

[1] Now we review the *Gamma distribution* and *Inverted Gamma distribution* as follows:

(1) A random variable X has a *Gamma distribution* with pdf:

$$p(x|\alpha, \beta) = \mathcal{G}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad \text{for } x > 0,\ \alpha > 0, \text{ and } \beta > 0.$$

(2) Define $Y = 1/X$. Then the new random variable Y has an *Inverted Gamma distribution* with density:

$$p(y|\alpha, \beta) = \mathcal{IG}(y|\alpha, \beta) = d\left(x^{-1}\right) p\left(x^{-1}|\alpha, \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\beta/y} \qquad \text{for } y > 0.$$

- Then *conditional posterior* for the regression coefficients is given by:

$$
\begin{aligned}
p(\beta|X,Y,\sigma^2) &\propto p(Y|X,\beta,\sigma^2)p(\beta) \\
&\propto \exp\left\{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0) - \frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\beta'(\Sigma_0^{-1}+\sigma^{-2}X'X)\beta - 2\beta'(\Sigma_0^{-1}\beta_0+\sigma^{-2}X'Y)\right]\right\} \\
&= \mathcal{N}(\beta_1,\Sigma_1).
\end{aligned}
$$

where

$$
\begin{aligned}
\beta_1 &= \Sigma_1\left(\Sigma_0^{-1}\beta_0+\sigma^{-2}X'Y\right) \\
\Sigma_1 &= \left(\Sigma_0^{-1}+\sigma^{-2}X'X\right)^{-1}.
\end{aligned}
$$

Note that σ^2 is known.

- The *conditional posterior* for variance is similarly

$$
\begin{aligned}
p(\sigma^2|X,Y,\beta) &\propto p(Y|X,\beta,\sigma^2)p(\sigma^2) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+T}{2}+1} \exp\left\{-\frac{\delta_0}{2\sigma^2} - \frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} \\
&= \mathcal{IG}\left(\frac{\nu_1}{2}, \frac{\delta_1}{2}\right).
\end{aligned}
$$

where

$$
\begin{aligned}
\nu_1 &= T + \nu_0 \\
\delta_1 &= (Y-X\beta)'(Y-X\beta) + \delta_0.
\end{aligned}
$$

Note that β is known.

4.2 Multivariate regression models

- Consider a *multivariate* normal model for a vector $\boldsymbol{Y}_t$:

$$\boldsymbol{Y}_t \sim \mathcal{N}(\boldsymbol{B}\boldsymbol{X}_t, \boldsymbol{\Sigma}) \quad \text{or} \quad \boldsymbol{Y}_t \sim \mathcal{N}(\boldsymbol{Z}_t\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

where $\boldsymbol{Z}_t = (\boldsymbol{X}_t' \otimes I)$, $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B})$, $\boldsymbol{Y}_t$ is a $k \times 1$ vector, $\boldsymbol{X}_t$ is a $r \times 1$ vector, $\boldsymbol{B}$ is a $k \times r$ matrix and $\boldsymbol{\Sigma}$ is a $k \times k$ covariance matrix.

- We assume the *inverse Wishart ($\mathcal{IW}$)* prior distribution for $\boldsymbol{\Sigma}$:

$$\begin{aligned}\boldsymbol{\Sigma} \sim \mathcal{IW}(m_0, \boldsymbol{\Psi}_0; k) &= \frac{|\boldsymbol{\Psi}_0|^{m_0/2}|\boldsymbol{\Sigma}|^{-(m_0+k+1)/2}\exp(-\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_0)/2)}{2^{m_0k/2}\Gamma_k(m_0/2)} \\ &\propto |\boldsymbol{\Sigma}|^{-(m_0+k+1)/2}\exp(-\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_0)/2)\end{aligned}$$

where $\Gamma_k(\cdot)$ is the multivariate gamma function.

- The *prior distribution* for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{A}_0) = -(2\pi)^{-k/2}|\boldsymbol{A}_0|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{A}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}$$
$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{A}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}.$$

- The *likelihood function* is given by:

$$p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{\beta})\right\}$$
$$\propto |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{\beta})(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{\beta})'\right)\right\}.$$

- Then the *conditional posterior* for $\boldsymbol{\beta}$ is given by:

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\Sigma}) &\propto p(\boldsymbol{Y}|\boldsymbol{\Sigma}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) \\
&\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (\boldsymbol{Y}_t - \boldsymbol{Z}_t \boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_t - \boldsymbol{Z}_t \boldsymbol{\beta}) - \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{A}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \\
&\propto \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_1)' \boldsymbol{A}_1^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_1) \right\} \\
&\propto \mathcal{N}(\boldsymbol{\beta}_1, \boldsymbol{A}_1),
\end{aligned}
$$

where

$$
\boldsymbol{\beta}_1 = \boldsymbol{A}_1 \left(\boldsymbol{A}_0^{-1} \boldsymbol{\beta}_0 + \sum_{t=1}^{T} \boldsymbol{Z}_t' \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}_t \right), \quad \boldsymbol{A}_1 = \left(\boldsymbol{A}_0^{-1} + \sum_{t=1}^{T} \boldsymbol{Z}_t' \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}_t \right)^{-1}.
$$

- The *condtional posteriors* for the variance is similarly straightforward:

$$
\begin{aligned}
p(\boldsymbol{\Sigma}|\boldsymbol{Y},\boldsymbol{\beta}) &\propto p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}) \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{m_0+k+1+T}{2}} \exp\left\{-\frac{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_0)+\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{t=1}^{T}(\boldsymbol{Y}_t-\boldsymbol{Z}_t\boldsymbol{\beta})(\boldsymbol{Y}_t-\boldsymbol{Z}_t\boldsymbol{\beta})'\right)}{2}\right\} \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{m_0+k+1+T}{2}} \exp\left\{-\frac{\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}[\boldsymbol{\Psi}_0+\sum_{t=1}^{T}(\boldsymbol{Y}_t-\boldsymbol{Z}_t\boldsymbol{\beta})(\boldsymbol{Y}_t-\boldsymbol{Z}_t\boldsymbol{\beta})']\}}{2}\right\} \\
&\propto \mathcal{IW}(m_1,\boldsymbol{\Psi}_1;k)
\end{aligned}
$$

where $m_1 = m_0 + T$ and $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_0 + \sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{\beta})(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{\beta})'$.

4.3 State space models

Consider next a linear state space model in which a scalar observation y_t, is generated as

$$
\begin{aligned}
y_t &= \boldsymbol{x}_t' \boldsymbol{\beta}_t + u_t, \quad u_t \sim \mathcal{N}(0, \sigma^2), \\
\boldsymbol{\beta}_t &= \boldsymbol{G} \boldsymbol{\beta}_{t-1} + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim \mathcal{N}_m(0, \boldsymbol{\Psi}), \quad 1 \leq t \leq n \\
\sigma^2 &\sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right), \quad \boldsymbol{\Psi} \sim \mathcal{IW}_m(\rho_0, R_0),
\end{aligned}
$$

where $\boldsymbol{\beta}_t$ is an $m \times 1$ state vector and $\boldsymbol{G}$ is assumed known.

Carter and Kohn (1994), Fruhwirth-Schnatter (1994), de Jong and Shephard (1995).

- The MCMC implementation for this model is based on distributions

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n | \boldsymbol{y}, \sigma^2, \boldsymbol{\Psi}; \quad \sigma^2 | \boldsymbol{y}, \{\boldsymbol{\beta}_t\}, \boldsymbol{\Psi}; \quad \boldsymbol{\Psi} | \boldsymbol{y}, \{\boldsymbol{\beta}_t\}.$$

- To see how the $\boldsymbol{\beta}_t$'s are sampled, write the joint distribution as

$$p(\boldsymbol{\beta}_n | \boldsymbol{y}, \sigma^2, \boldsymbol{\Psi}) \times p(\boldsymbol{\beta}_{n-1} | \boldsymbol{y}, \boldsymbol{\beta}_n, \sigma^2, \boldsymbol{\Psi}) \times \cdots \times p(\boldsymbol{\beta}_1 | \boldsymbol{y}, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n, \sigma^2, \boldsymbol{\Psi}),$$

where, on letting $\boldsymbol{\beta}^s = (\boldsymbol{\beta}_s, \ldots, \boldsymbol{\beta}_n)$, $\boldsymbol{Y}_s = (y_1, \ldots, y_s)$ and $\boldsymbol{Y}^s = (y_s, \ldots, y_n)$ for $s < n$, the typical term is

$$\begin{aligned} p(\boldsymbol{\beta}_t | \boldsymbol{y}, \sigma^2, \boldsymbol{\beta}^{t+1}, \boldsymbol{\Psi}) &= p(\boldsymbol{\beta}_t | \boldsymbol{Y}_t, \sigma^2, \boldsymbol{\beta}_{t+1}, \boldsymbol{\Psi}) \\ &\propto p(\boldsymbol{\beta}_t | \boldsymbol{Y}_t, \sigma^2, \boldsymbol{\Psi}) p(\boldsymbol{\beta}_{t+1} | \boldsymbol{\beta}_t, \sigma^2, \boldsymbol{\Psi}), \end{aligned}$$

due to the fact that $(\boldsymbol{Y}^{t+1}, \boldsymbol{\beta}^{t+2})$ is independent of $\boldsymbol{\beta}_t$ given $(\boldsymbol{\beta}_{t+1}, \sigma^2, \boldsymbol{\Psi})$.

- The first density on the right hand side is Gaussian with moments given by the Kalman filter recursions, i.e.,

$$\boldsymbol{\beta}_t \mid \boldsymbol{Y}_t, \sigma^2, \boldsymbol{\Psi} \sim \mathcal{N}(\boldsymbol{\beta}_{t|t}, \boldsymbol{R}_{t|t}).$$

- The second density is Gaussian with moments $\boldsymbol{G}\boldsymbol{\beta}_t$, and $\boldsymbol{\Psi}$, i.e.,

$$\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t, \sigma^2, \boldsymbol{\Psi} \sim \mathcal{N}(\boldsymbol{G}\boldsymbol{\beta}_t, \boldsymbol{\Psi}).$$

- Proceeding the problem as if the observation equation

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{G}\boldsymbol{\beta}_t + \boldsymbol{u}_{t+1}, \quad \boldsymbol{u}_{t+1} \sim \mathcal{N}(0, \boldsymbol{\Psi})$$

and the transition equation

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t|t} + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim \mathcal{N}(0, \boldsymbol{R}_{t|t}).$$

- The joint distribution for $\boldsymbol{\beta}_{t+1}$ and $\boldsymbol{\beta}_t$, given $\boldsymbol{Y}_t, \sigma^2, \boldsymbol{\Psi}$, is

$$\begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix} \Bigg| \boldsymbol{Y}_t, \sigma^2, \boldsymbol{\Psi} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{G}\boldsymbol{\beta}_{t|t} \\ \boldsymbol{\beta}_{t|t} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G}\boldsymbol{R}_{t|t}\boldsymbol{G}' + \boldsymbol{\Psi} & \boldsymbol{G}\boldsymbol{R}_{t|t} \\ \boldsymbol{R}_{t|t}\boldsymbol{G}' & \boldsymbol{R}_{t|t} \end{bmatrix} \right)$$

- By Gaussian Conditional Theorem, the moments of $p(\boldsymbol{\beta}_t|\boldsymbol{y}, \boldsymbol{\beta}^{t+1}, \sigma^2, \boldsymbol{\Psi}) \sim \mathcal{N}(\boldsymbol{\beta}_{t|t,\boldsymbol{\beta}_{t+1}}, \boldsymbol{R}_{t|t,\boldsymbol{\beta}_{t+1}})$ can be easily derived, i.e.,

$$\boldsymbol{\beta}_{t|t,\boldsymbol{\beta}_{t+1}} = \boldsymbol{\beta}_{t|t} + \boldsymbol{M}_t(\boldsymbol{\beta}_{t+1} - \boldsymbol{G}\boldsymbol{\beta}_{t|t})$$

$$\boldsymbol{R}_t = \boldsymbol{R}_{t|t} - \boldsymbol{M}_t\boldsymbol{R}_{t+1|t}\boldsymbol{M}_t',$$

  where $\boldsymbol{M}_t = \boldsymbol{R}_{t|t}\boldsymbol{G}'\boldsymbol{R}_{t+1|t}^{-1} = \boldsymbol{R}_{t|t}\boldsymbol{G}'(\boldsymbol{G}\boldsymbol{R}_{t|t}\boldsymbol{G}' + \boldsymbol{\Psi})^{-1}$.

- Then, the joint distribution $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n|\boldsymbol{y}, \sigma^2, \boldsymbol{\Psi}$ can be sampled by the method of forward-filtering-backward-sampling (FFBS).

Algorithm 6 (Gaussian state space)**.**

1. *Sample state vectors using forward-filtering-backward-sampling (FFBS) algorithm.*

   (a) *Kalman filter. Calculate for* $t = 1, 2, \ldots, n$

$$\widehat{\boldsymbol{\beta}}_{t|t-1} = \boldsymbol{G}\widehat{\boldsymbol{\beta}}_{t-1|t-1}, \qquad \boldsymbol{R}_{t|t-1} = \boldsymbol{G}\boldsymbol{R}_{t-1|t-1}\boldsymbol{G}' + \boldsymbol{\Psi},$$

$$f_{t|t-1} = \boldsymbol{x}_t'\boldsymbol{R}_{t|t-1}\boldsymbol{x}_t + \sigma^2, \qquad \boldsymbol{K}_t = \boldsymbol{R}_{t|t-1}\boldsymbol{x}_t f_{t|t-1}^{-1},$$

$$\widehat{\boldsymbol{\beta}}_{t|t} = \widehat{\boldsymbol{\beta}}_{t|t-1} + \boldsymbol{K}_t(y_t - \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_{t|t-1}), \qquad \boldsymbol{R}_{t|t} = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{x}_t')\boldsymbol{R}_{t|t-1},$$

$$\boldsymbol{M}_t = \boldsymbol{R}_{t|t}\boldsymbol{G}'\boldsymbol{R}_{t+1|t}^{-1}, \qquad \boldsymbol{R}_t = \boldsymbol{R}_{t|t} - \boldsymbol{M}_t\boldsymbol{R}_{t+1|t}\boldsymbol{M}_t'.$$

   Store $\widehat{\boldsymbol{\beta}}_{t|t}$*;* $\boldsymbol{M}_t$*;* $\boldsymbol{R}_t$*.*

   (b) *Sample* $\boldsymbol{\beta}_n \sim \mathcal{N}_m(\widehat{\boldsymbol{\beta}}_{n|n}, \boldsymbol{R}_{n|n})$*.*

(c) Sample for $t = n-1, n-2, \ldots, 1$

$$\boldsymbol{\beta}_t \sim \mathcal{N}_m(\widehat{\boldsymbol{\beta}}_{t|t,\boldsymbol{\beta}_{t+1}}, \boldsymbol{R}_t), \quad \widehat{\boldsymbol{\beta}}_{t|t,\boldsymbol{\beta}_{t+1}} = \widehat{\boldsymbol{\beta}}_{t|t} + \boldsymbol{M}_t(\boldsymbol{\beta}_{t+1} - \boldsymbol{G}\widehat{\boldsymbol{\beta}}_{t|t}).$$

2. Sample

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\nu_0 + n}{2}, \frac{\delta_0 + \sum_{t=1}^n (y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_t)^2}{2}\right\}.$$

3. Sample

$$\boldsymbol{\Psi} \sim \mathcal{IW}_m\left\{\rho_0 + n, \left[\boldsymbol{R}_0^{-1} + \sum_{t=1}^n (\boldsymbol{\beta}_t - \boldsymbol{G}\boldsymbol{\beta}_{t-1})(\boldsymbol{\beta}_t - \boldsymbol{G}\boldsymbol{\beta}_{t-1})'\right]^{-1}\right\}.$$

4. Goto 1.

4.4 Markov switching models

The general Markov switching model is described as

$$
\begin{aligned}
y_t \mid \boldsymbol{Y}_{t-1}, S_t = k, \boldsymbol{\theta} &\sim f(y_t \mid \boldsymbol{Y}_{t-1}, \boldsymbol{\theta}_k), \quad k = 1, \ldots, M \\
S_t \mid S_{t-1}, \boldsymbol{P} &\sim \text{Markov}(\boldsymbol{P}, \pi_1), \\
\boldsymbol{\theta} &\sim p(\boldsymbol{\theta}), \quad \text{such as } \mathcal{N} \text{ and } \mathcal{IG} \\
\boldsymbol{p}_i &\sim \text{Dirichlet}(\alpha_{i1}, \ldots, \alpha_{iM}), \quad i \leq M
\end{aligned}
$$

where $S_t \in \{1, \ldots, M\}$ is an unobservable random variable which evolves according to a Markov process with transition matrix $\boldsymbol{P} = \{p_{ij}\}$, with $p_{ij} = \Pr(S_t = j | S_{t-1} = i)$ and initial distribution π_1 at $t = 1$, and $\boldsymbol{p}_i$ is the i-th column of $\boldsymbol{P}$ that is assumed to have a Dirichlet prior distribution with parameters $(\alpha_{1i}, \ldots, \alpha_{Mi})$.

By CH theorem, the conditional distribution $p(\boldsymbol{S}_n, \boldsymbol{\theta}, \boldsymbol{P}|\boldsymbol{y})$ can be fully determined by the conditional distributions

$$\{S_t\}_{t=1}^n|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{P}, \quad \boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{S}_n, \boldsymbol{P}, \quad \{\boldsymbol{p}_i\}_{i=1}^M|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{S}_n$$

- Chib (1996) suggests an efficient algorithm to sample the states jointly in the first set of blocks from $p(\boldsymbol{S}_n|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{P}) = \Pr(\boldsymbol{S}_n|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{P}) =$

  $$\Pr(S_n|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{P}) \times \Pr(S_{n-1}|\boldsymbol{y}, S_n, \boldsymbol{\theta}, \boldsymbol{P}) \times \cdots \times \Pr(S_1|\boldsymbol{y}, S_2, \ldots, S_n, \boldsymbol{\theta}, \boldsymbol{P}),$$

  with typical term (for $k = 1, \ldots, M$ and $\boldsymbol{S}^{t+1} = \{S_{t+1}, \ldots, S_n\}$)

  $$\Pr(S_t = k|\boldsymbol{y}, \boldsymbol{S}^{t+1}, \boldsymbol{\theta}, \boldsymbol{P}) = \Pr(S_t = k|\boldsymbol{Y}_t, S_{t+1}, \boldsymbol{\theta}, \boldsymbol{P})$$
  $$= \frac{\Pr(S_t = k|\boldsymbol{Y}_t, \boldsymbol{\theta}, \boldsymbol{P}) \Pr(S_{t+1}|S_t = k, \boldsymbol{\theta}, \boldsymbol{P})}{\sum_{\ell=1}^M \Pr(S_t = \ell|\boldsymbol{Y}_t, \boldsymbol{\theta}, \boldsymbol{P}) \Pr(S_{t+1}|S_t = \ell, \boldsymbol{\theta}, \boldsymbol{P})},$$

- The first probability is the filtered probability of $S_t = k$,

$$\Pr(S_t = k|\boldsymbol{Y}_t, \boldsymbol{\theta}, \boldsymbol{P}) = \frac{\Pr(S_t = k|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P}) \times f(y_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}_k, \boldsymbol{P})}{\sum_{\ell=1}^{M} \Pr(S_t = \ell|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P}) \times f(y_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}_\ell, \boldsymbol{P})}$$

  with the forecasting probability of $S_t = k$,

$$\begin{aligned}
&\Pr(S_t = k|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P}) \\
&= \sum_{\ell=1}^{M} \Pr(S_t = k|S_{t-1} = \ell, \boldsymbol{P}) \times \Pr(S_{t-1} = \ell|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P}).
\end{aligned}$$

- Similar to the state space models, the sampling of $\boldsymbol{S}_n$ is also achieved by the forward-filtering-backward-sampling (FFBS) algorithm.

– **Forward filtering:** Running the Hamilton filter to get the forecasting and filtered probabilities of $\boldsymbol{S}_n$.

– **Backward sampling:** Sampling $\boldsymbol{S}_n$ in the backward pass.

(1) Firstly, simulating S_n from the probability mass function $\Pr(S_n|\boldsymbol{Y}_n, \boldsymbol{\theta}, \boldsymbol{P})$.

(2) Secondly, simulating S_t from the probability mass function $\Pr(S_t = k|\boldsymbol{y}, S^{t+1}, \boldsymbol{\theta}, \boldsymbol{P})$ recursively for $t = n-1, n-2, \ldots, 1$.

- Given the simulated $\boldsymbol{S}_n$, the data separates into M pieces and the simulation of $\boldsymbol{\theta}_k$ is from the full conditional distribution, $k = 1, \ldots, M$,

$$\pi(\boldsymbol{\theta}_k|\boldsymbol{y}, \boldsymbol{S}_n, \boldsymbol{\theta}_{-k}, \boldsymbol{P}) \propto p(\theta_k) \prod_{t:S_t=k} f(y_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P})$$

- Finally, the last distribution depends simply on $\boldsymbol{S}_n$ with each column $\boldsymbol{p}_i$ of $\boldsymbol{P}$ independently an updated Dirichlet distribution:

$$\boldsymbol{p}_i|\boldsymbol{S}_n \sim \text{Dirichlet}(\alpha_{i1} + n_{i1}, \ldots, \alpha_{iM} + n_{iM})$$

where n_{ik} is the total number of one-step transitions from state i to state k in the vector $\boldsymbol{S}_n$. If only two regimes are considered, the Dirichlet distribution is reduced to a Beta distribution.

Algorithm 7 (Markov switching model).

1. Calculate and store from $t = 1, \ldots, n$, $\Pr(S_t | \boldsymbol{Y}_t, \boldsymbol{\theta}, \boldsymbol{P})$.

2. Sample $S_n \sim \Pr(S_n | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{P})$.

3. For $t = n-1, n-2, \ldots, 1$*, sample* $S_t \sim \Pr(S_t | \boldsymbol{y}, \boldsymbol{S}^{t+1}, \boldsymbol{\theta}, \boldsymbol{P})$.

4. For $k = 1, \ldots, M$*, sample* $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k) \prod_{t:S_t=k} f(y_t | \boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P})$.

5. For $i = 1, \ldots, M$*, sample* $\boldsymbol{p}_i \sim \text{Dirichlet}(\alpha_{i1} + n_{i1}, \ldots, \alpha_{iM} + n_{iM})$.

6. Goto 1.

5 MCMC for univariate SV models

5.1 Single-move MCMC, JPR (1994)

- Jacquier et al. (1994) consider the basic SV model

$$y_t = \sqrt{V_t}\varepsilon_t$$

$$\log V_t = \nu + \phi \log V_{t-1} + \sigma_\eta \eta_t$$

- The only difficulty in this model is sampling from $p(V|\nu, \phi, \sigma_\eta^2, y)$. This distribution is not a *recognizable distribution*, and due to its high dimension, a direct application of MH is not recommended.

- Jacquier et al. (1994) suggest to break the T-dimensional distribution $p(V|\nu, \phi, \sigma_\eta^2, y)$ into T number of 1-dimensional distributions, i.e.,

$$
\begin{aligned}
p(V_t|V_{t-1}, V_{t+1}, \theta, y_t) &= \frac{p(y_t, V_{t+1}, V_t, V_{t-1}|\theta)}{p(y_t, V_{t+1}, V_{t-1}|\theta)} \\
&\propto p(y_t|V_t)p(V_{t+1}|V_t, \theta)p(V_t|V_{t-1}, \theta)p(V_{t-1}|\theta) \\
&\propto p(y_t|V_t)p(V_{t+1}|V_t, \theta)p(V_t|V_{t-1}, \theta) \\
&\propto \frac{1}{V_t^{0.5}} \exp\left\{-\frac{y_t^2}{2V_t}\right\} \times \exp\left\{-\frac{(\log V_{t+1} - \nu - \phi \log V_t)^2}{2\sigma_\eta^2}\right\} \\
&\quad \times \frac{1}{V_t} \exp\left\{-\frac{(\log V_t - \nu - \phi \log V_{t-1})^2}{2\sigma_\eta^2}\right\} \\
&\propto \frac{1}{V_t^{0.5}} \exp\left\{-\frac{y_t^2}{2V_t}\right\} \times \frac{1}{V_t} \exp\left\{-\frac{(\log V_t - \mu_t)^2}{2\sigma^2}\right\},
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_t &= (\nu(1-\phi) + \phi(\log V_{t+1} + \log V_{t-1}))/(1+\phi^2) \\
\sigma^2 &= \sigma_\eta^2/(1+\phi^2).
\end{aligned}
$$

- As this joint density is not recognizable, they use an *independence Metropolis-Hastings* algorithm to sample from it.

  1. JPR choose an *inverse Gamma* proposal density $q(V_t)$ motivated by the observation that the first term in the posterior is an *inverse Gamma* and the second log-normal term can be approximated (particularly in the tails) by a *suitable chosen inverse Gamma*.
  2. The resulting blanket

$$q(V_t|\cdot) \propto V_t^{-(\Phi+1)} e^{-\Theta_t/V_t} \rightarrow \mathcal{IG}(\Phi, \Theta_t),$$

  where

$$\Phi = \Phi_1 + \Phi_{LN} + 1, \quad \Phi_1 = -0.5, \quad \Phi_{LN} = (1 - 2e^{\sigma^2})/(1 - e^{\sigma^2})$$

$$\Theta_t = \Theta_{1,t} + \Theta_{LN,t}, \quad \Theta_{1,t} = y_t^2/2, \quad \Theta_{LN,t} = (\Phi_{LN} - 0.5)\exp(\mu_t + 0.5\sigma^2).$$

3. Given that the *inverse gamma distribution* bounds the tails of the true conditional density, the algorithm is *geometrically convergent*.

- Thus, the hybrid MCMC algorithm for estimating the stochastic volatility requires the following steps: given

1. Draw $\nu^{(g+1)}, \phi^{(g+1)} \mid \sigma_\eta^{2(g)}, V^{(g)} \sim \mathcal{N}$
2. Draw $\sigma_\eta^{2(g+1)} \mid \nu^{(g+1)}, \phi^{(g+1)}, V^{(g)} \sim \mathcal{IG}$
3. Draw $V_t^{(g+1)}$ with the *MH* algorithm.
   (a) Draw $V_t^{(g+1)} \sim q\left(V_t | V_{t-1}^{(g+1)}, V_t^{(g)}, V_{t+1}^{(g)}, \theta^{(g+1)}\right)$ for $t = 1, \ldots, T$.
   (b) Accept $V_t^{(g+1)}$ with probability $\alpha\left(V_t^{(g)}, V_t^{(g+1)}\right)$, where

$$\alpha\left(V_t^{(g)}, V_t^{(g+1)}\right) = \min\left\{\frac{\pi(V_t^{(g+1)})q(V_t^{(g)})}{\pi(V_t^{(g)})q(V_t^{(g+1)})}, 1\right\}.$$

5.2 Single-move MCMC, Shephard (1996)

- Shephard (1996) and Kim et al. (1998) proposed another *single-move MCMC sampler* for another formulation of the basic SV model

$$y_t = \exp(h_t/2)\varepsilon_t$$

$$h_{t+1} = \nu + \phi h_t + \sigma_\eta \eta_t,$$

- Other than Jacquier et al. (1994), they break the joint density $p(h|y,\theta)$ into T univariate conditional distributions $p(h_t|h_{t+1}, h_{t-1}, y_t, \theta)$,

$$p(h_t|h_{t+1}, h_{t-1}, y_t, \theta) \propto p(h_t|h_{t+1}, h_{t-1}, \theta)p(y_t|h_t, \theta), \quad t = 1, \ldots, T.$$

  We sample this density by developing a *simple accept/reject procedure*.

- It can be shown that

$$p(h_t|h_{t+1}, h_{t-1}, \theta) \propto \exp\left\{-\frac{(h_{t+1} - \nu - \phi h_t)^2 + (h_t - \nu - \phi h_{t-1})^2}{2\sigma_\eta^2}\right\}$$
$$= \phi_N(h_t|h_t^*, \upsilon^2)$$

where

$$\upsilon^2 = \frac{\sigma_\eta^2}{1+\phi^2}, \quad h_t^* = \mu + \frac{\phi\{(h_{t-1} - \mu) + (h_{t+1} - \mu)\}}{(1+\phi^2)},$$

and we let

$$\mu := \frac{\nu}{1-\phi}.$$

- Next we note that $\exp(-h_t)$ is convex function (凸函数) and can be bounded by a function linear in h_t. Thus, we have $\exp(-h_t) \geq \exp(-h_t^*) - \exp(-h_t^*)(h_t - h_t^*)$.

- Let $\log p(y_t|h_t,\theta) = \text{const} + \log p^*(y_t,h_t,\theta)$. Then

$$\begin{aligned}\log p^*(y_t,h_t,\theta) &= -\frac{1}{2}h_t - \frac{y_t^2}{2}\{\exp(-h_t)\} \\ &\leq -\frac{1}{2}h_t - \frac{y_t^2}{2}\{\exp(-h_t^*)(1+h_t^*) - h_t\exp(-h_t^*)\} \\ &= \log g^*(y_t,h_t,\theta,h_t^*).\end{aligned}$$

- Hence we have

$$p(h_t|h_{t+1},h_{t-1},\theta)p^*(y_t,h_t,\theta) \leq \phi_N(h_t|h_t^*,\upsilon^2)g^*(y_t,h_t,\theta,h_t^*)$$

$$p(h_t|h_{t+1},h_{t-1},\theta)p(y_t|h_t,\theta) \leq \phi_N(h_t|\mu_t,\upsilon^2).$$

where

$$\mu_t = h_t^* + \frac{\upsilon^2}{2}[y_t^2\exp(-h_t^*) - 1].$$

- Hence with these results, the accept-reject procedure to sample h_t from $p(h_t|h_{t+1}, h_{t-1}, y_t, \theta)$ can now be implemented.

  1. Propose a value h_t from $N(\mu_t, \upsilon^2)$.
  2. Accept this value with probability p^*/g^*; if rejected return to the first step and make a new proposal.

In Kim et al. (1998), they suggest this MCMC algorithm given as follows

$$
\begin{aligned}
p(\nu, \phi|\sigma_\eta^2, h, y) &\sim \mathcal{N} \\
p(\sigma_\eta^2|\nu, \phi, h, y) &\sim \mathcal{IG} \\
p(h_t|h_{\setminus t}, \nu, \phi, \sigma_\eta, y) &\sim Accept/Reject, \quad \text{for } t = 1, \ldots, T.
\end{aligned}
$$

5.3 Multi-move or blocking MCMC

The true log-chi-square density with one degree of freedom, $x \sim \log \chi_1^2$, is

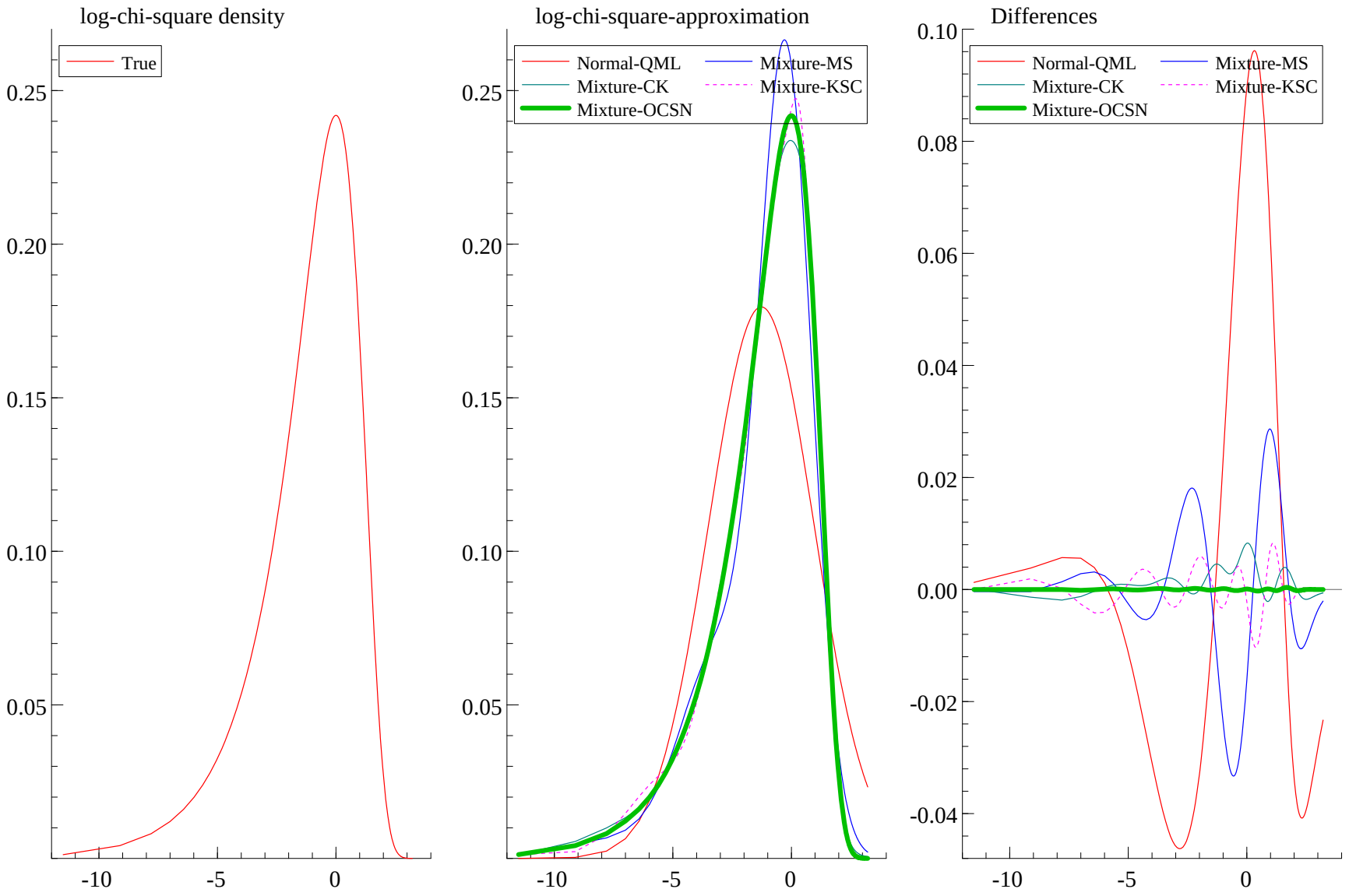
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{x - \exp(x)}{2}\right\}.$$

There are many studies introducing the idea of accurately approximating the $\log \chi_1^2$ distribution by a matched mixture of normal distributions, e.g.

- Carter and Kohn (1997), $K = 5$.
- Mahieu and Schotman (1998), $K = 3$.
- Kim et al. (1998), $K = 7$.
- Omori et al. (2007), $K = 10$. (Best!)

	$B=3$, MS (1998)				$B=5$, CK (1997)		
b	ω_b	μ_b	Q_b	b	ω_b	μ_b	Q_b
1	0.0500	−6.4818	9.00120	1	0.1300	−4.6300	8.7500
2	0.2500	−3.0461	2.46647	2	0.1600	−2.8700	1.9500
3	0.7000	−0.2172	1.22147	3	0.2300	−1.4400	0.8800
				4	0.2200	−0.3300	0.4500
				5	0.2500	0.7600	0.4100
	$B=7$, KSC (1998)				$B=10$, OCSN (2007)		
1	0.00730	−11.40039	5.79596	1	0.00115	−14.65000	7.33342
2	0.00002	−9.83726	5.17950	2	0.01575	−8.68384	4.16591
3	0.10556	−5.24321	2.61369	3	0.05591	−5.55246	2.54498
4	0.25750	−2.35859	1.26261	4	0.12047	−3.46788	1.57469
5	0.34001	−0.65098	0.64009	5	0.18842	−1.97278	0.98583
6	0.24566	0.52478	0.34023	6	0.22715	−0.85173	0.62699
7	0.04395	1.50746	0.16735	7	0.20674	0.02266	0.40611
				8	0.13057	0.73504	0.26768
				9	0.04775	1.34744	0.17788
				10	0.00609	1.92677	0.11265

注：本表中混合分布参数值根据 Mahieu & Schotman (1998) (MS)、Carter & Kohn (1997) (CK)、Kim et al. (1998) (KSC) 和 Omori et al. (2007) (OCSN) 的相关结果获得。


log-chi-square density
True
0.25
0.20
0.15
0.10
0.05
-10
-5
0
log-chi-square-approximation
Normal-QML
Mixture-MS
Mixture-CK
Mixture-KSC
Mixture-OCSN
0.25
0.20
0.15
0.10
0.05
-10
-5
0
Differences
Normal-QML
Mixture-MS
Mixture-CK
Mixture-KSC
Mixture-OCSN
0.10
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-10
-5
0

- Kim, Shephard and Chib (1998) consider alternative MCMC method to estimate the model, allowing *block updating*.

- They also use the log-form of the transformed SV model

$$y_t^* = h_t + \log(\varepsilon_t^2)$$

$$h_t = \nu + \phi h_{t-1} + \sigma_\eta \eta_t.$$

- They argue that $\xi_t = \log(\varepsilon_t^2)$ can be easily approximated by a 7-component mixture of normals:

$$p(\xi_t) \approx \sum_{j=1}^{7} q_j \phi(x_t | \mu_j, \sigma_j^2)$$

where $\phi(x|\mu, \sigma^2)$ is normal distributed with mean μ and variance σ^2.

- Formally, this requires the addition of a latent state variable, s_t, such that

$$\xi_t | s_t = j \sim \mathcal{N}(\mu_j, \sigma_j^2), \qquad \Pr(s_t = j) = q_j$$

- Then the transformed model is rewritten as

$$y_t^* = h_t + \mu_{s_t} + u_t, \quad u_t \sim N(0, \sigma_{s_t}^2)$$

$$h_t = \nu + \phi h_{t-1} + \sigma_\eta \eta_t$$

$$\Pr(s_t = j) = q_j \quad \text{and} \quad \sum_{j=1}^{7} q_j = 1.$$

- In this transformed model, the posterior is $p\left(\nu, \phi, \sigma_\eta^2, h, s | y\right)$.

- Conditional on the indicators, the model is a linear, normal state space model and the Kalman recursions deliver $p(h|\nu, \phi, \sigma_\eta, s, y)$.

- The multi-period posterior

$$p(s|\nu, \phi, \sigma_\eta, h, y) = \Pi_{t=1}^{T} \Pr(s_t|\theta, h_t, y_t)$$

  is transformed into T one-period posterior

$$\Pr(s_t = j|\theta, h_t, y_t) \propto p(y_t|s_t = j, \theta, h_t) \Pr(s_t = j)$$

  Thus, we can draw the discrete variable $\{s_t\}_{t=1}^{T}$ using the inverse transform method introduced previously.

The MCMC algorithm is now

$$p(\nu, \phi | \sigma_\eta^2, h, s, y) \sim \mathcal{N}$$

$$p(\sigma_\eta^2 | \nu, \phi, h, s, y) \sim \mathcal{IG}$$

$$p(h | \nu, \phi, \sigma_\eta, s, y) \sim FFBS$$

$$p(s | \nu, \phi, \sigma_\eta, h, y) \sim Multinomial.$$

- Further details of the algorithm and the exact conditional posteriors are given in Kim, Shephard and Chib (1998).

- This algorithm generates a Markov Chain with very low autocorrelations. Due to the low autocorrelations, the algorithm is often referred to as a rapidly converging algorithm.

6 MCMC for Time-varying Parameter VAR Models

Note: The time-varying parameter VAR (vector autoregressive) model has been widely used in economics over the past two decades, in areas such as economic policy analysis, financial risk analysis, and network connectivity analysis. However, the estimation of this model remains a frontier issue actively explored by the academic community, primarily due to the urgent need for rapid and efficient estimation methods amidst the expansion of big data and high-dimensional models.

Reference:

Nakajima, J. (2011), "Time-Varying Parameter VAR Model with Stochastic Volatility - An Overview of Methodology and Empirical Applications" Monetary and Economic Studies, 29, 107-142.

Structural form representation

We begin with a basic structural VAR model defined as

$$Ay_t = F_1 y_{t-1} + \ldots + F_p y_{t-p} + u_t, \quad u_t \sim N(0, \Sigma\Sigma), \tag{2}$$

where $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_N\}$ and A is a lower-triangular matrix with the diagonal elements being ones. Rewrite it as the reduced-form model

$$y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + A^{-1}\Sigma\varepsilon_t, \quad \varepsilon_t \sim N(0, I)$$

Furthermore, defining $\boldsymbol{X}_t = I_N \otimes (y'_{t-1}, \ldots, y'_{t-p})$, the model can be rewritten as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + A^{-1}\Sigma\varepsilon_t.$$

Time-varying parameter model extension

Consider the time varying parameter (TVP) VAR model specified by

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta}_t + A_t^{-1} \Sigma_t \varepsilon_t. \tag{3}$$

where $\boldsymbol{X}_t = I_N \otimes (y'_{t-1}, \ldots, y'_{t-p})$, $\boldsymbol{\beta}_t = \mathrm{vec}((B_{1t}, B_{2t}, \ldots, B_{pt})')$,

$$A_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ a_{21,t} & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ a_{N1,t} & \cdots & a_{NN-1,t} & 1 \end{bmatrix}, \quad \Sigma_t = \begin{bmatrix} \sigma_{1t} & 0 & \cdots & 0 \\ 0 & \sigma_{2t} & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{Nt} \end{bmatrix}.$$

Following Primiceri (2005), let

$$\boldsymbol{\alpha}_t = (a_{21,t}, a_{31,t}, a_{32,t}, a_{41,t}, \ldots, a_{N,N-1,t})',$$

$$\boldsymbol{h}_t = (h_{1t}, \ldots, h_{Nt})', \quad \text{with } h_{jt} = \log \sigma_{jt}^2 \text{ for } j = 1, \ldots, N.$$

It is assumed that the parameters in (3) follow a random walk process

$$\begin{aligned}\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \boldsymbol{u}_{\beta t},\\ \boldsymbol{\alpha}_{t+1} &= \boldsymbol{\alpha}_t + \boldsymbol{u}_{\alpha t},\\ \boldsymbol{h}_{t+1} &= \boldsymbol{h}_t + \boldsymbol{u}_{ht},\end{aligned} \qquad \begin{pmatrix} \varepsilon_t \\ \boldsymbol{u}_{\beta t} \\ \boldsymbol{u}_{\alpha t} \\ \boldsymbol{u}_{ht} \end{pmatrix} \sim N\left(0, \begin{pmatrix} I & O & O & O \\ O & \Sigma_\beta & O & O \\ O & O & \Sigma_\alpha & O \\ O & O & O & \Sigma_h \end{pmatrix}\right)$$

where $\Sigma_h = \mathrm{diag}\{\sigma_{h1}^2, \ldots, \sigma_{hN}^2\}$ is a diagonal matrix.

Estimation method

Given the data $\boldsymbol{y}$, we draw samples from the posterior distribution, $\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{h}, \Sigma_\beta, \Sigma_\alpha, \Sigma_h | \boldsymbol{y})$, by the following MCMC algorithm:

1. Initialize $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{h}$ and Σ_β, Σ_α, Σ_h.
2. Sample $\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{h}, \Sigma_\beta, \boldsymbol{y}$ by FFBS (forward filtering and backward sampling).
3. Sample $\Sigma_\beta|\boldsymbol{\beta}$ from Inverse Wishart posterior.
4. Sample $\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{h}, \Sigma_\alpha, \boldsymbol{y}$ by FFBS.
5. Sample $\Sigma_\alpha|\boldsymbol{\alpha}$ from Inverse Wishart posterior.
6. Sample $\boldsymbol{h}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \Sigma_h, \boldsymbol{y}$ by single-move or multi-move sampling.
7. Sample $\sigma_{hi}^2|\boldsymbol{h}$ from Inverse Gamma posterior, $i = 1, \ldots, N$.
8. Go to 2.

Applications

- The original intention is to capture the transmission path of monetary policies, see for example Cogley and Sargent (2005), Primiceri (2005), Koop et al. (2009) and Nakajima (2011).
- In a study of network spillovers among financial institutions, Geraci and Gnabo (2018) demonstrate the utility of TVP-VAR models for a system of four sectors.
- Similarly, Ciccarelli and Rebucci (2007) use a TVP-VAR to study contagion and interdependence among exchange rates.
- In psychology, the TVP-VAR is used to model emotion dynamics and has been proposed for the study of networks in psychopathology (精神病理学网络) (see Bringmann et al. (2018) and references therein).