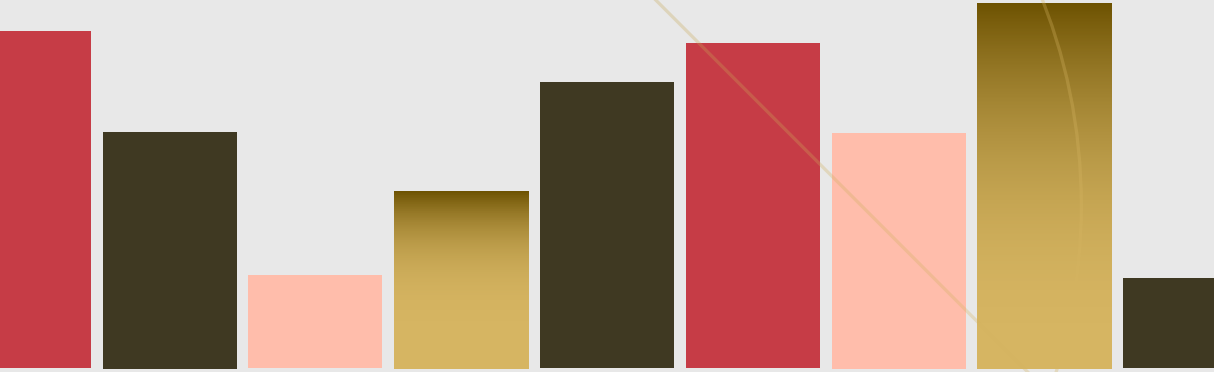

Accept-Reject Metropolis-Hastings sampling and Marginal likelihood estimation

- Chib and Jeliazkov (2005)

Statistica Neerlandica

Research Idea

A new method for estimating the marginal likelihood when the simulation from the posterior distribution of the model parameters is by accept-reject Metropolis-Hastings Algorithm. We use ARMH method when we don't know the normalizing constant of the proposal density (used in MH). If sampling is done by Gibbs sampler or MH algorithm then Chib (1995) or Chib and Jeliazkov (2001) will be used respectively.

Problem Set up

$$m(y) = \int f(y|\theta)\,\pi(\theta)\,d\theta$$

$m(y)$: marginal likelihood (used in calculation of Bayes factor for model comparison)

$f(y|\theta)$: Sampling density of y

$\pi(\theta)$: prior density of θ

$$m(y) = \frac{f(y|\theta)\,\pi(\theta)}{\pi(\theta|y)}$$ → Basic marginal likelihood identity (found using Bayes rule)

Taking logs on both sides gives

$$\log m(y) = \underbrace{\log f(y|\theta^*) + \log \pi(\theta^*)}_{\text{available using direct calculation}} - \underbrace{\log \pi(\theta^*|y)}_{\text{find estimate of posterior density evaluated at } \theta^*}$$

For estimation efficiency, we select θ^* at some high density region.

Let the parameters are grouped into B blocks.

$\theta = (\theta_1, \theta_2, \dots \theta_B)$ with $\theta_k \in \mathbb{R}^d$, $k = 1, 2 \dots B$.

Let $\psi_i = (\theta_1, \theta_2, \dots, \theta_i)$, $\psi^{i+1} = (\theta_{i+1}, \theta_{i+2}, \dots, \theta_B)$

Therefore, we can write the posterior density as

$$\pi(\theta_1^*, \dots, \theta_B^* | y) = \pi(\theta_1^* | y)\, \pi(\theta_2^* | y, \theta_1^*) \cdots \pi(\theta_B^* | y, \theta_1^* \cdots \theta_{B-1}^*)$$

$$= \prod_{i=1}^{B} \pi(\theta_i^* | y, \theta_1^*, \dots, \theta_{i-1}^*)$$

$$= \prod_{i=1}^{B} \pi(\theta_i^* | y, \psi_{i-1}^*)$$

Side definitions

Let x_i's are i.i.d. θ is the concerned parameter

* Sufficient statistics: a statistic $T = f(X_1, X_2, \dots X_n)$ (function of data) is sufficient for θ if

$f(X_1, X_2 \dots X_n, \theta | T(x) = t)$ is not a function of θ.
(conditional probability are the function in terms of things conditioned on)

t contains all the relevent information from the

data for θ. t has extracted everything from the θ regarding t.

* Factorization Theorem [For calculating sufficient statistics]
  - If $T(x)$ is sufficient for θ if
    $f(x_1, x_2, \dots x_n, \theta) = g_\theta(T(x))\, h(x_1, x_2 \dots x_n)$

Example : Let $\theta = \mu$ & $x_i \sim N(\mu, \sigma^2)$ i.id

$$f(x_1, x_2, \dots x_n, \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 + n\mu^2 - 2\mu\sum_{i=1}^{n} x_i\right)\right\}$$

$$= \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right\}}_{h(x_1, x_2 \dots x_n)} \underbrace{\exp\left\{-\frac{1}{2\sigma^2}\left\{n\mu^2 - 2\mu\sum_{i=1}^{n} x_i\right\}\right\}}_{g_\theta(T(x))}$$

(terms without μ) (terms with μ in it)

There, the only function of data in $g_\theta(T(x))$ is $\sum_{i=1}^{n} x_i$.
Thus, it is a sufficient statistics

* Rao - Blackwell Theorem. - If $h(x)$ is an unbiased estimator of θ i.e $E(h(x)) = \theta$, then $E[h(x) \mid T(x) = t] = g(t)$ is an unbiased estimator of θ and has minimum variance. [t is sufficient statistics]
In other words, if we have an unbiased estimator a which is a function of sufficient statistics t, then it is a unique, unbiased and most efficient estimator

Case 1: full conditionals are known

Chib (1995) proposed $\pi(\theta_i^* \mid y, \psi_{i-1}^*)$ by Rao-Blackwellization

$$\pi(\theta_i^* \mid y, \psi_{i-1}^*) = \int \underbrace{\pi(\theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1})}_{\text{full conditional that we know for gibbs sampling}} \underbrace{\pi(\psi^{i+1} \mid y, \psi_{i-1}^*)}_{\text{draw } \psi^{i+1} \text{ from here}} d\psi^{i+1}$$

$$= G^{-1} \sum_{g=1}^{G} \pi(\theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1,g})$$

Case 2: full conditionals are not known

Chib and Jeliazkov (2001) use local reversibility to show

$$\pi(\theta_i^* \mid y, \psi_{i-1}^*) = \frac{E_1\{\alpha_{MH}(\theta_i, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1}) q(\theta_i, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1})\}}{E_2\{\alpha_{MH}(\theta_i^*, \theta_i \mid y, \psi_{i-1}^*, \psi^{i+1})\}}$$

E_1 is expectation under $\pi(\psi^i \mid y, \psi_{i-1}^*)$

E_2 is expectation under $\pi(\psi^{i+1} \mid y, \psi_i^*) q(\theta_i^*, \theta_i \mid y, \psi_{i-1}^*, \psi^{i+1})$

ARMH algorithm

We use it when full conditionals are not known and normalizing constant of proposal density is not known.

We are interested in sampling the target density $\pi(x)$.
$\pi(x) = f(x)/K$ where K is unknown [marginal distribution]

There is a pdf $h(\cdot)$ available for sampling. Now for any c, it is not necessary that $ch(\cdot)$ is greater than $f(x)$ $\forall$ c [as was the case in Accept-Reject sampling].

let $C = \{x : f(x) < ch(x)\}$
C^c is the complement of C.

Given $x^{(n)} = x$, we have to sample $x^{(n+1)} = y$.

density of the new r.v $\leftarrow [q(y) = P\left(y \mid U \le \frac{f(z)}{ch(z)}\right)$

It will serve as proposal density for MH step.

y being sampled [y is sampled from $h(\cdot)$] $= \dfrac{P\left(U \le \frac{f(z)}{ch(z)} \mid z = y\right) \times h(y)}{P\left(U \le \frac{f(z)}{ch(z)}\right)}$] Bayes rule [like how we calculate posterior distribution]

let $P\left(U \le \frac{f(z)}{ch(z)}\right) = \frac{f(z)}{ch(z)}$ be some constant d.

$$q(y) = \frac{P\left(U \le \frac{f(z)}{c\,h(z)} \mid z = y\right) \times h(y)}{d}$$

but $P\left(U \le \frac{f(z)}{c\,h(z)} \mid z = y\right) = \min\left\{\frac{f(z)}{c\,h(z)}, 1\right\}$

$$q(y) = \left[\min\left\{\frac{f(y)}{c\,h(z)}, 1\right\} * h(y)\right] \frac{1}{d}$$

Therefore, $q(y) = f(y)/cd$ if $y \in C$

$q(y) = \frac{h(y)}{d}$ if $y \in C^c$

We writing $q(y)$ instead of $q(x,y)$ as y is generated independent of x.

$q(y)$ is the candidate generating density. As we don't know the normalizing constant for the proposal density, we are drawing observations from the AR step and then will pass on to MH step.

Now there are 4 possibilities for x & y which will go into MH step.

a) $x, y \in C$

b) $x \in C$ & $y \in C^c$

c) $x \in C^c$ & $y \in C$

d) $x, y \in C^c$

Now in MH step, we have to find moving probability $\alpha(x,y)$ which satisfy reversibility

Thus, we will derive $\alpha(x,y)$ in all the four cases.

$$f(x)\, q(y)\, \alpha(x,y) = f(y)\, q(x)\, \alpha(y,x)$$

$f(x)$: target density

$q(y)$: transitional density [generally we chose in MH step]. we rather created it using AR step.

Case 1 : $x \in C,\ y \in C$

replacing for $q(\cdot) = f(\cdot)/cd$

$$\frac{f(x)}{cd}\,\frac{f(y)}{cd}\,\alpha(x,y) = \frac{f(y)}{cd}\,\frac{f(x)}{cd}\,\alpha(y,x)$$

$$\alpha(x,y) = \alpha(y,x) = 1$$

Hence, we directly select y in this case with probability 1. Kind of selecting it in AR process

Case 2 : $x \in C,\ y \notin C$ or $x \notin C$ & $y \in C$

replacing for $q(\cdot)$

We will show for $x \in C$ & $y \notin C$

$$f(x) \frac{f(y)}{c\,d} \alpha(x,y) < f(y) \frac{h(x)}{d} \alpha(y,x)$$

because $f(y) > c\,h(y) \Rightarrow h(y) < \frac{f(y)}{c}$

multiplying both sides by $f(x)/d$

$$\frac{f(x)\,h(y)}{d} < \frac{f(y)\,f(x)}{c\,d} \Rightarrow f(x)\,q(y) < f(y)\,q(x)$$

$\therefore \alpha(x,y) = 1, \alpha(y,x) = c\frac{h(x)}{f(x)}$

Similarly, if $x \notin C$ and $y \in C$

$$\alpha(x,y) = \frac{c\,h(y)}{f(y)}, \quad \alpha(y,x) = 1$$

In this case, AR sample has to go through another step of MH.

In case $x \in C$ & $y \notin C$, $\alpha(x,y) = 1$.

In case $x \notin C$ & $y \in C$, $\alpha(x,y) < 1$ as $f($

Case 3: $x \notin C$ and $y \notin C$

Case 3.1: too many transitions from x to y as compared from y to x.

$$f(x) \frac{h(y)}{d} \alpha(x,y) > f(y) \frac{h(x)}{d} \alpha(y,x)$$

$$\therefore \alpha(y,x) = 1 \quad , \alpha(x,y) = \frac{f(y)\,h(x)}{f(x)\,h(y)}$$

Case 3.2 : There are too little transitions from x to y as compared from y to x

$$\therefore \alpha(y,x) = \frac{f(x)\,h(y)}{f(y)\,h(x)} \quad , \alpha(x,y) = 1$$

In general : $\alpha(x,y) = \min\left\{\frac{f(y)\,h(x)}{f(x)\,h(y)}, 1\right\}$

or

$$\alpha(y,x) = \min\left\{\frac{f(x)\,h(y)}{f(y)\,h(x)}, 1\right\}$$

In the cases where $x \in C$, the probability of move to y is 1 i.e we selected them in accept-reject step only.

To summarize, that's how the steps are followed

1) Let $C_1 = \{f(x) < ch(x)\}$ & $C_2 = \{f(y) < ch(y)\}$

2) Generate U from $(0,1)$
   - if $C_1 = 1$, then select y and $\alpha_{xy} = 1$
   - if $C_1 = 0$ and $C_2 = 1$, then $\alpha_{xy} = ch(x)/f(x)$
   - if $C_1 = 0$ and $C_2 = 0$, then $\alpha_{xy} = \min\left\{\frac{f(y)\,h(x)}{f(x)\,h(y)}, 1\right\}$

3) if $u \leq \alpha$
   - return y
4) if $u > \alpha$
   - return x

When $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$ is the target density which is the posterior distribution. Then

1) Generate $\theta' \sim h(\theta|y)$, accept θ' with probability

$$\alpha_{AR}(\theta'|y) = \min\left\{1, \frac{f(y|\theta')\pi(\theta')}{ch(\theta'|y)}\right\}$$

$$= \min\left\{1, \frac{f(y|\theta')}{ch(\theta'|y)}\right\}$$

Continue until θ' is accepted

2) Given current value θ & proposal value θ'

a) if $\theta \in C$, set $\alpha_{MH}(\theta, \theta'|y) = 1$

b) if $\theta \notin C$ and $\theta' \in C$, set $\alpha_{MH}(\theta, \theta'|y) = \frac{ch(\theta|y)}{f(y|\theta)\pi(\theta)}$

c) if $\theta \notin C$ and $\theta' \notin C$, set $\alpha_{MH}(\theta, \theta'|y)$

$$= \min\left\{1, \frac{f(y|\theta')\pi(\theta')h(\theta|y)}{f(y|\theta)\pi(\theta)h(\theta'|y)}\right\}$$

Return θ' with probability $\alpha_{MH}(\theta, \theta'|y)$ else θ.

Draw from AR step have density

$$q(\theta|y) = \frac{\alpha_{AR}(\theta|y)\, h(\theta|y)}{\int \alpha_{AR}(\theta|y)\, h(\theta|y)\, d\theta}$$

which serves as proposal density for MH step the denominator which is the normalizing constant is not available analytically and that's why we added AR step.

Proposed Approach (one block case)

Chib and Jeliazkov (2001) use reversibility of the Markov chain to obtain estimate for marginal distribution.

$$\pi(\theta^*|y) = \frac{\int \alpha_{MH}(\theta, \theta^*|y)\, q(\theta, \theta^*|y)\, \pi(\theta|y)\, d\theta}{\int \alpha_{MH}(\theta^*, \theta|y)\, q(\theta^*, \theta|y)\, d\theta}$$

Now we will use the above estimate but will substitute for $q(\theta, \theta^*|y)$ (numerator) and $q(\theta^*, \theta|y)$ (denominator) which was proposed density in MH case with $q(\theta^*|y)$ and $q(\theta|y)$ (not dependent on previous value θ) which is the proposed density derived from AR step.

Denoting $q(\theta^*|y)$ to correct the context

$$q(\theta^*|y) = \text{Prob}\left(\theta^* \mid y, U \le \frac{f(\theta^*|y)}{ch(\theta^*|y)}\right)$$

$$q(\theta^*|y) = \frac{\text{Prob}\left(U \le \frac{f(\theta^*|y)}{ch(\theta^*|y)} \mid \theta^*, y\right) h(\theta^*|y)}{\int \text{Prob}\left(U \le \frac{f(\theta^*|y)}{ch(\theta^*|y)}\right) h(\theta^*|y)\, d\theta^*}$$

$$q(\theta^*|y) = \frac{\alpha_{AR}(\theta^*|y)\, h(\theta^*|y)}{d}$$

Similarly

$$q(\theta|y) = \frac{\alpha_{AR}(\theta|y)\, h(\theta|y)}{d}$$

Substituting in Chib and Jeliazkov (2001) result

$$\pi(\theta^*|y) = \frac{\int \alpha_{MH}(\theta, \theta^*|y)\, q(\theta^*|y)\, \pi(\theta|y)\, d\theta}{\int \alpha_{MH}(\theta^*, \theta|y)\, q(\theta|y)\, d\theta}$$

$$\pi(\theta^*|y) = \frac{\int \alpha_{MH}(\theta, \theta^*|y)\, d^{-1}\alpha_{AR}(\theta^*|y)\, h(\theta^*|y)\, \pi(\theta|y)\, d\theta}{\int \alpha_{MH}(\theta^*, \theta|y)\, q(\theta|y)\, d\theta}$$

If we restrict $\theta^* \in C$, then the above expression will simplify as $\alpha_{MH}(\theta^*, \theta|y) = 1$

$$\pi(\theta^*|y) = \frac{d^{-1}\alpha_{AR}(\theta^*|y)\, h(\theta^*|y) \int \alpha_{MH}(\theta, \theta^*|y)\, \pi(\theta|y)\, d\theta}{\int q(\theta|y)\, d\theta}$$

$$\pi(\theta^*|y) = \frac{d^{-1}\alpha_{AR}(\theta^*|y)\, h(\theta^*|y) \int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\, d\theta}{\int q(\theta|y)\, d\theta}$$

as $\theta^* \in C$, $\alpha_{AR}(\theta^*|y) = \frac{f(y|\theta^*)\pi(\theta^*)}{ch(\theta^*|y)}$

$$\pi(\theta^*|y) = \frac{f(y|\theta^*)\pi(\theta^*)\, h(\theta^*|y)}{dch(\theta^*|y)} \int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\, d\theta \quad \left(\int q(\theta|y)\, d\theta \hookrightarrow = 1\right)$$

$$\pi(\theta^*|y) = \frac{f(y|\theta^*)\ \pi(\theta^*) \int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\, d\theta}{dc}$$

$$\pi(\theta^*|y) = \frac{f(y|\theta^*)\ \pi(\theta^*) \int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\, d\theta}{c\int \alpha_{AR}(\theta|y)\, h(\theta|y)\, d\theta}$$

if we substitute the above value in

$$m(y) = \frac{f(y|\theta^*)\pi(\theta^*)}{\pi(\theta^*|y)}$$

Then,

$$m(y) = \frac{c\int \alpha_{AR}(\theta|y)\, h(\theta|y)\, d\theta}{\int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\, d\theta}$$

We can find an estimate using the above result.

$$\hat{m}(y) = c \; \frac{\frac{1}{J}\sum_{j=1}^{J} \alpha_{AR}(\theta^{j}|y)}{\frac{1}{G}\sum_{g=1}^{G} \alpha_{MH}(\theta^{(g)}, \theta^{*}|y)}$$

where numerator $\theta^{(j)}$ is from $h(\theta|y)$] AR step
denominator $\theta^{(g)}$ is from $\pi(\theta|y)$] MH step where only those θ sampled from $h(\theta|y)$ are included which got accepted.

Note: draws from ARMH are closer to i.i.d than the ones sampled from MH. Hence, the variance of ARMH draws will be smaller.

Proposed Approach (Multi Block case)

$h(\theta_i \mid y, \psi_{i-1}, \psi^{i+1})$ - AR proposal density.

Let $C_i = \{\theta_i : f(y \mid \psi_{i-1}, \psi^{i+1})\, \pi(\theta_i \mid \psi_{i-1}, \psi^{i+1}) \leq c_i(\psi_{i-1}, \psi^{i+1})\, h(\theta_i|y, \psi_{i-1}, \psi^{i+1})\}$

MH proposal density

$$q(\theta_i \mid y, \psi^{*}_{i-1}, \psi^{i+1}) = \frac{\alpha_{AR}(\theta_i^{*}|y, \psi^{*}_{i-1}, \psi^{i+1})\, h(\theta_i^{*}|y, \psi^{*}_{i-1}, \psi^{i+1})}{d(y, \psi^{*}_{i-1}, \psi^{i+1})}$$

$d(y, \psi^{*}_{i-1}, \psi^{i+1})$: normalizing constant which is unknown.

Now using Chib and Jeliazkov (2001)

$$\pi(\theta_i^* \mid y, \psi_{i-1}^*) = \frac{E_1\{\alpha_{MH}(\theta_i, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1}) \; q(\theta_i, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1})\}}{E_2\{\alpha_{MH}(\theta_i^*, \theta_i \mid y, \psi_{i-1}^*, \psi^{i+1}\}}$$

Substituting for $q(\theta_i, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1})$

$$\pi(\theta_i^* \mid y, \psi_{i-1}^*) = \frac{E_1\{\alpha_{MH}(\theta_i, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1}) \; \alpha_{AR}(\theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1}) \, h(\theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1})\}}{E_2\{\alpha_{MH}(\theta_i^*, \theta_i \mid y, \psi_{i-1}^*, \psi^{i+1}) \, \alpha_{AR}(\theta_i \mid y, \psi_{i-1}^*, \psi^{i+1})\}}$$

E_1 is expectation with respect to $\pi(\psi^i \mid y, \psi_{i-1}^*)$
E_2 is expectation with respect to $\pi(\psi^{i+1} \mid y, \psi_i^*)$
$h(\theta_i \mid y, \psi_{i-1}^*, \psi^{i+1})$

We will not simplify $\pi(\theta_i^* \mid y, \psi_{i-1}^*)$ as we did in the one block case as it might lead to loss of efficiency.
In order to sample from the above expression, the following steps are followed.

1) Set $\psi_{i-1} = \psi_{i-1}^*$ and sample the set of full conditional distributions $\pi(\theta_k \mid y, \theta_{-k})$ $k = i, \ldots, B$
Generated values be $(\theta_i^{(g)}, \ldots \theta_B^{(g)})$ [Normal

simulations using ARMH)

2) In the conditioning set, fix θ_i at θ_i^* to produce $\psi_i^* = \{\psi_{i-1}^*, \theta^*\}$ and sample the remaining distribution $\pi(\theta_k \mid y, \theta_{-k})$, $k = i+1, \ldots, B$

Generate $\{\theta_{i+1}^{(j)} \ldots \theta_B^{(j)}\}$

at each step draw $\theta_i^{(j)} \sim h(\theta_i \mid y, \psi_{i-1}^*, \psi^{i+1,(j)})$

3) Estimate the marginal density as

$$\frac{\frac{1}{G}\sum_{g=1}^{G} \alpha_{MH}(\theta_i^{(g)}, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1,g})\, \alpha_{AR}(\theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1,g})\, h(\theta_i^* \mid y, \psi_{i-1}\, \psi^{i+1,g})}{\frac{1}{G}\sum_{j=1}^{G} \alpha_{MH}(\theta_i^*, \theta_i^{(j)} \mid y, \psi_{i-1}^*, \psi^{i+1,j})\, \alpha_{AR}(\theta_i^{(j)} \mid y, \psi_{i-1}^*, \psi^{i+1,j})}$$

4) $\log \hat{m}(y) = \log(y \mid \theta^*) + \log \pi(\theta^*) - \sum_{i=1}^{B} \log \hat{\pi}(\theta_i \mid y, \theta_1^* \ldots \theta_{i-1}^*)$

is the estimated value.