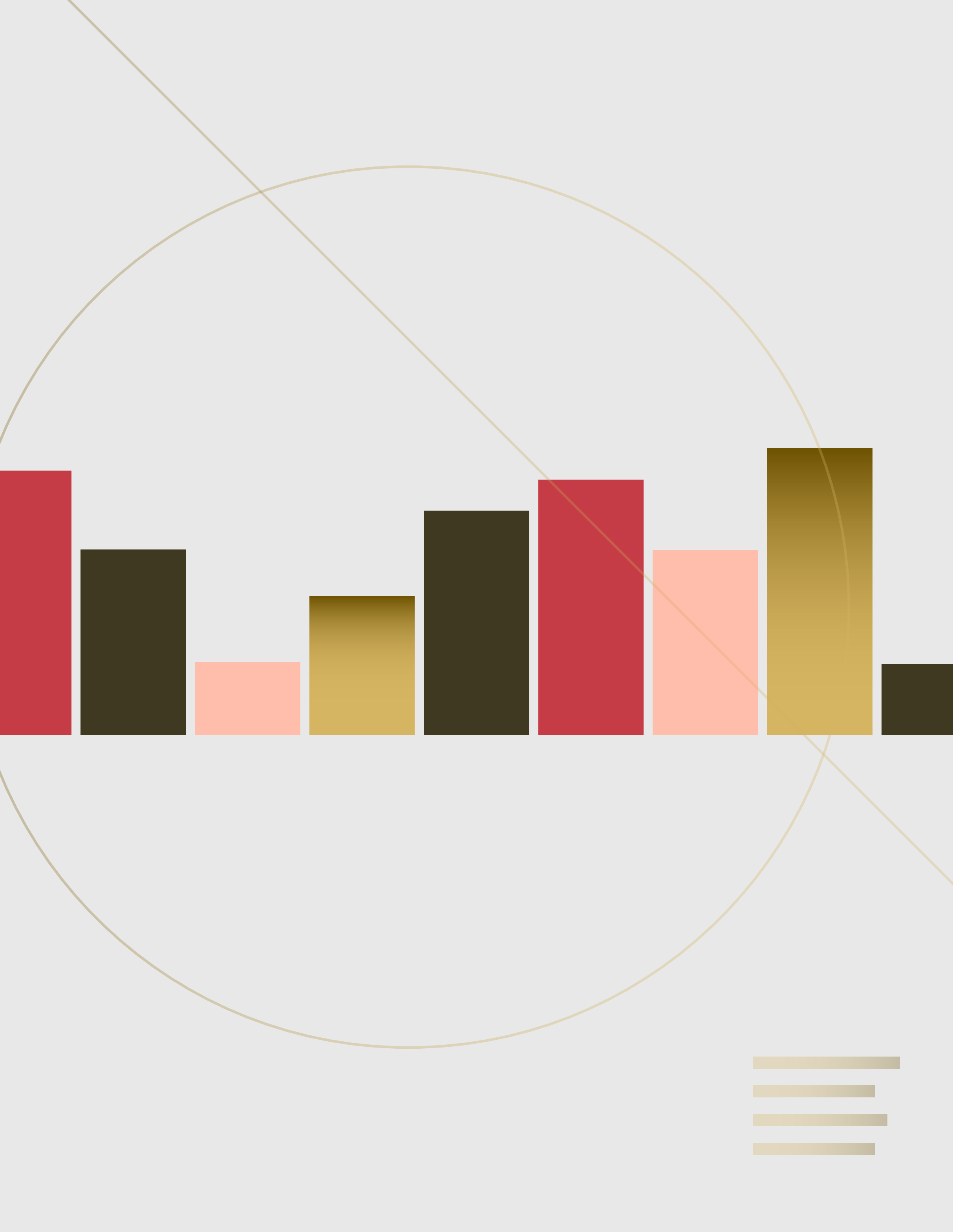

Marginal likelihood from the Metropolis - Hastings output
- Chib and Jeliazkov (2001)
Journal of American Statistical Association

Research Question

The marginal likelihood goes into the calculation of Bayes factor which is used to compare different models. But marginal likelihood is obtained by integrating the sampling density $f(y|M_1, \theta_1)$ with respect to the prior distribution of the parameters, Hence, the posterior MCMC output from the simulation cannot be used directly to estimate the marginal likelihood. The focus is on those problems where full conditionals cannot be calculated else we can use Chibb (1995).

Basic Marginal likelihood identity

$$\underbrace{\pi(\theta_l | y, M_l)}_{\text{posterior distribution of model } l} = \frac{\overbrace{f(y|M_l, \theta_l)}^{\text{conditional density of } y \text{ or likelihood function for model } l} \; \overbrace{\pi(\theta_l | M_l)}^{\text{Prior distribution of } \theta_l \text{ for model } l.}}{\underbrace{m(y|M_l)}_{\text{Marginal likelihood of model } l.}}$$

$$m(y \mid M_\ell) = \frac{f(y \mid M_\ell, \theta_\ell)\, \pi(\theta_\ell \mid M_\ell)}{\pi(\theta_\ell \mid y, M_\ell)} \quad \text{- BMI}$$

we can solve for LHS if we take logs on both sides and evaluate at θ^*

$$\log m(y \mid M_\ell) = \log(f(y \mid M_\ell, \theta_\ell^*)) + \log \pi(\theta_\ell^* \mid M_\ell) - \log \pi(\theta_\ell^* \mid y, M_\ell)$$

The first two terms are known [Prior and likelihood]. We just need to estimate posterior density of θ_ℓ at θ_ℓ^*.

It's recommended to take θ_ℓ^* as high density point for efficiency.

Review of Metropolis-Hastings Algorithm

Aim is to have reversible kernel s.t

$$f(x)\, p(x,y) = f(y)\, p(y,x)$$

in other words, the probability of going from state X to state Y is same as the probability from state Y to state X.

When we have full conditionals, they represent the transitional probability from 1 state to another as $P(\theta_1^{s+1} \mid \theta_2^s \cdots \theta_k^s, y)$ where $\theta_2^s \cdots \theta_k^s$ were sampled based on θ_1^s and so on. Thus it is a transition.

Though, we don't have full conditionals now, we have to come up with transitional probability by ourself

Define a kernal $q(\cdot,\cdot)$ which satisfy reversibility.

$$\underbrace{f(x)}_{\text{distribution of } x} q(x,y) = f(y)\, q(y,x)$$

[we will consider posterior density of x & not prior]

We can show that using reversibility, we can sample from posterior distribution. (Chib & Greenberg (1995))

But, it is very difficult to find $q(\cdot,\cdot)$ which satisfy reversibility. Hence, multiply both sides by a function $\alpha(x,y)$ which gives weight to transition probability $q(x,y)$ to balance the equation.

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

If LHS > RHS, then $\alpha(y,x) = 1$ & $\alpha(x,y) = \dfrac{f(y)\, q(y,x)}{f(x)\, q(x,y)} < 1$

If LHS < RHS, then $\alpha(x,y) = 1$ & $\alpha(y,x) = \dfrac{f(x)\, q(x,y)}{f(y)\, q(y,x)} < 1$

$\alpha(x,y)\, q(x,y) \rightarrow$ if the present state of the process is x, generate a value y from $q(x,y)$ and make the move with probability $\alpha(x,y)$.

$\therefore$ transition kernal $= \alpha(x,y)\, q(x,y)$

combines continuous kernel and PMF $\alpha(x,y)$

[analogous to full conditionals in Gibbs sampler]

$q(x,y)$ = probability that next state is y given the current state is x. Both x & y are some states of the parameter θ. It will go to those states of θ where density is high with higher likelihood.
It is like sampling for A from $P(A|B)$. the only difference is that in each iteration, A becomes B and a new A is sampled.

$\alpha(x,y)$ = probability with which the transition from x to y will take place.

Derivation of the approach

Suppose posterior density $\pi(\theta|y) \propto \pi(\theta) f(y|\theta)$ is sampled using Metropolis - Hastings algorithm and the goal is to estimate $\pi(\theta^*|y)$ which will go in BMI.

One block Sampling

Let $q(\theta,\theta'|y)$ denote the proposal density for trasition from θ to θ' and is allowed to depend on data y.

Let

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

probability of move [accepting the proposed value θ']

$$\therefore \underbrace{p(\theta, \theta'|y)}_{\text{transitional probability}} = \alpha(\theta, \theta'|y)\, q(\theta, \theta'|y)$$

Using the reversibility condition (for θ^*)

$$\underbrace{\pi(\theta|y)}_{\text{posterior distribution of } \theta} p(\theta, \theta^*|y) = \underbrace{\pi(\theta^*|y)}_{\text{posterior distribution of } \theta^* \text{ that we want to calculate}} p(\theta^*, \theta|y)$$

$$\pi(\theta|y)\, \alpha(\theta, \theta^*|y)\, q(\theta, \theta^*|y) = \pi(\theta^*|y)\, \alpha(\theta^*, \theta|y)\, q(\theta^*, \theta|y)$$

integrating both sides by θ
[accounting for or summing over all θ]

$$\int \pi(\theta|y)\, \alpha(\theta, \theta^*|y)\, q(\theta, \theta^*|y)\, d\theta = \int \pi(\theta^*|y)\, \alpha(\theta^*, \theta|y)\, q(\theta^*, \theta|y)\, d\theta$$

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

We can write the above expression as

$$\pi(\theta^* | y) = \frac{E_1\{\alpha(\theta, \theta^* | y)\, q(\theta, \theta^* | y)\}}{E_2\{\alpha(\theta^*, \theta | y\}}$$

$$\int \underbrace{\alpha(\theta, \theta^* | y)\, q(\theta, \theta^* | y)}_{x}\, \underbrace{\pi(\theta | y)}_{f(x)\ [\text{posterior}]}\, d\theta = E_1(x)$$

$$\int \underbrace{\alpha(\theta^*, \theta | y)}_{x}\, q\underbrace{(\theta^*, \theta)}_{f(x)\ [\text{proposed density}]}\, d\theta = E_2(x)$$

Estimator is $\hat{\pi}(\theta^* | y) = \dfrac{\frac{1}{M} \sum_{g=1}^{M} \alpha(\theta^{(g)}, \theta^* | y)\, q(\theta^{(g)}, \theta^* | y)}{\frac{1}{J} \sum_{j=1}^{J} \alpha(\theta^*, \theta^{(j)} | y)}$

$\theta^{(g)}$ is sampled draws from posterior distribution $\pi(\theta | y)$ [It was the $f(x)$ while calculating $E_1(x)$]

$\theta^{(j)}$ is sampled draws from $q(\theta^*, \theta | y)$ [It was $f(x)$ while calculating $E_2(x)$]

Substituting $\hat{\pi}(\theta^* | y)$ in log BMI.

$$\underbrace{\log \hat{m}(y)}_{\text{final estimator}} = \log f(y | \theta^*) + \log \pi(\theta^*) - \log \hat{\pi}(\theta^* | y)$$

Note : we can keep sample draws M and J [for numerator and denominator] same.

Two Block Sampling

we assume multiple latent variables block
$z = \{z_1, z_2 \ldots z_n\}$

θ_1 and θ_2 are Parameters block
we don't know full conditional for θ_1

$q(\theta_1, \theta_1' \mid y, \theta_2, z)$ is the proposal density for the transition from θ_1 to θ_1'.

$\alpha(\theta_1, \theta_1' \mid y, \theta_2, z)$ = Probability of move

$$= \min \left\{ 1, \frac{f(y \mid \theta_1', \theta_2, z)\, \pi(\theta_1', \theta_2)\; q(\theta_1', \theta_1 \mid y, \theta_2, z)}{\underbrace{f(y \mid \theta_1, \theta_2, z)\; \pi(\theta_1, \theta_2)}_{}\; q(\theta_1, \theta_1' \mid y, \theta_2, z)} \right\}$$

posterior density only written as likelihood and prior distribution (marginal in the denominator got cancelled out)

$$m(y) = \frac{f(y \mid \theta_1^*, \theta_2^*)\, \pi(\theta_1^*, \theta_2^*)}{\pi(\theta_1^*, \theta_2^* \mid y)}$$

we have to estimate $\pi(\theta_1^*, \theta_2^* \mid y)$

$$\pi(\theta_1^*, \theta_2^* \mid y) = \pi(\theta_1^* \mid y)\, \pi(\theta_2^* \mid y, \theta_1^*)$$

$$p(\theta_1, \theta_1^* | y, \theta_2, z) = \alpha(\theta_1, \theta_1^* | y, \theta_2, z)\; q(\theta_1, \theta_1^* | y, \theta_2, z)$$

by reversibility condition

$$p(\theta_1, \theta_1^* | y, \theta_2, z)\, \pi(\theta_1 | y, \theta_2, z) = \pi(\theta_1^* | y, \theta_2, z)\, p(\theta_1^*, \theta_1 | y, z, \theta_2)$$

as we want $\pi(\theta_1^* | y)$ only from the equation, we will do some manipulation

multiply both sides by $\pi(\theta_2, z | y)$ & integrate over θ_2, θ_1 and z.

$$\iiint p(\theta_1, \theta_1^* | y, \theta_2, z)\, \pi(\theta_1 | y, \theta_2, z)\, \pi(\theta_2, z | y)\, d\theta_1\, d\theta_2\, dz$$
$$= \iiint p(\theta_1^*, \theta_1 | y, z, \theta_2)\, \pi(\theta_1^* | y, \theta_2, z)\, \pi(\theta_2, z | y)\, d\theta_1\, d\theta_2\, dz$$

$$\pi(\theta_1^* | y, \theta_2, z)\, \pi(\theta_2, z | y) = \pi(\theta_1^* | y)\, \pi(\theta_2, z | y, \theta_1^*)$$

$$\iiint p(\theta_1, \theta_1^* | y, \theta_2, z)\, \pi(\theta_1 | y, \theta_2, z)\, \pi(\theta_2, z | y)\, d\theta_1\, d\theta_2\, dz$$
$$= \iiint p(\theta_1^*, \theta_1 | y, z, \theta_2)\, \pi(\theta_1^* | y)\, \pi(\theta_2, z | y, \theta_1^*)\, d\theta_1\, d\theta_2\, dz$$

$$\pi(\theta_1 | y, \theta_2, z)\, \pi(\theta_2, z | y) = \pi(\theta_1, \theta_2, z | y)$$

$$\pi(\theta_1^* | y) = \frac{\iiint p(\theta_1, \theta_1^* | y, \theta_2, z)\, \pi(\theta_1, \theta_2, z | y)\, d\theta_1\, d\theta_2\, dz}{\iiint p(\theta_1^*, \theta_1 | y, z, \theta_2)\, \pi(\theta_2, z | y, \theta_1^*)\, d\theta_1\, d\theta_2\, dz}$$

$$\pi(\theta_1^* | y) = \frac{\iiint \alpha(\theta_1, \theta_1^* | y, \theta_2, z)\, q_2(\theta_1, \theta_1^* | y, \theta_2, z)\, \pi(\theta_1, \theta_2, z | y)\, d\theta_1\, d\theta_2\, dz}{\iiint \alpha(\theta_1^*, \theta_1 | y, \theta_2, z)\, q_2(\theta_1^*, \theta_1 | y, \theta_2, z)\, \pi(\theta_2, z | y, \theta_1^*)\, d\theta_1\, d\theta_2\, dz}$$

$$\pi(\theta_1^*|y) = \frac{E_1\{\alpha(\theta_1, \theta_1^*|y, \theta_2, z)\, q(\theta_1, \theta_1^*|y, \theta_2, z)\}}{E_2\{\alpha(\theta_1^*, \theta_1|y, \theta_2, z)}$$

$$\iiint \underbrace{\alpha(\theta_1, \theta_1^*|y, \theta_2, z)\, q_2(\theta_1, \theta_1^*|y, \theta_2, z)}_{x} \underbrace{\pi(\theta_1, \theta_2, z|y)}_{f(x)\ [\text{posterior}]} d\theta_1 d\theta_2 dz = E_1(x)$$

$$\iiint \underbrace{\alpha(\theta_1^*, \theta_1|y, \theta_2, z)}_{x} \underbrace{q_2(\theta_1^*, \theta_1|y, \theta_2, z)\, \pi(\theta_2, z|y, \theta_1^*)}_{f(x)} d\theta_1 d\theta_2 dz = E_2(x)$$

For numerator, we take draws $\{\theta_1^{(g)}, \theta_2^{(g)}, z^{(g)}\}_{g=1}^{M}$ from the full run as the output of M-H algorithm is from the posterior density.

For denominator, $\pi(\theta_2, z|y, \theta_1^*)$ is conditioned on θ_1^*, we have to continue MCMC simulations for additional J iterations from the full conditional densities

$$\pi(\theta_2|y, z, \theta_1^*) \quad \& \quad \pi(z|y, \theta_1^*, \theta_2)$$

[can use gibbs sampling as well]

Or use $\pi(\theta_2|y, z, \theta_1^*)$ instead of $p(\theta_1^*, \theta_2|y, z)$ in Metropolis-Hastings algorithm.

Then generate $\theta_1^{(j)} \sim q(\theta_1^*, \theta_1|y, \theta_2^{(j)}, z^{(j)})$

$\therefore$ we get a triple $(\theta_1^{(j)}, \theta_2^{(j)}, z^{(j)})$ sampled from $q_2(\theta_1^*, \theta_1|\theta_2, y, z)\, \pi(\theta_2, z|\theta_1^*, y)$

The estimator will look like

$$\hat{\pi}(\theta_1^* | y) = \frac{\frac{1}{M} \sum_{g=1}^{M} \alpha(\theta_1^{(g)}, \theta_1^* | y, \theta_2^{(g)}, z^{(g)}) \, q(\theta_1^{(g)}, \theta_1^* | y, \theta_2^{(g)}, z^{(g)})}{\frac{1}{J} \sum_{j=1}^{J} \alpha(\theta_1^*, \theta_1^{(j)} | y, \theta_2^{(j)}, z^{(j)})}$$

$$m(y) = \frac{f(y | \theta_1^*, \theta_2^*) \, \pi(\theta_1^*, \theta_2^*)}{\pi(\theta_1^*, \theta_2^* | y)}$$

$$m(y) = \frac{f(y | \theta_1^*, \theta_2^*) \, \pi(\theta_1^*, \theta_2^*)}{\pi(\theta_1^* | y) \, \pi(\theta_2^* | y, \theta_1^*)}$$

$$\log m(y) = \log f(y | \theta_1^*, \theta_2^*) + \log \pi(\theta_1^*, \theta_2^*) - \log \pi(\theta_1^* | y) - \log \pi(\theta_2^* | y, \theta_1^*)$$

we have calculated $\hat{\pi}(\theta_1^* | y)$ above
we need to calculate $\pi(\theta_2^* | y, \theta_1^*)$
As we assumed that we have full conditional for θ_2, we can use

$$\hat{\pi}(\theta_2^* | y, \theta_1^*) = \frac{1}{J} \sum_{j=1}^{J} \pi(\theta_2^* | y, \theta_1^*, z^{(j)})$$

where $z^{(j)}$ was sampled above

$$\therefore \log \hat{m}(y) = \log f(y | \theta^*) + \log \pi(\theta^*) - \log \hat{\pi}(\theta_1^* | y) - \log \hat{\pi}(\theta_2^* | y, \theta_1^*)$$

Multiple Block Sampling

No latent variable z

B blocks $= \{\theta_1, \theta_2 \cdots \theta_B\}$

posterior density : $\pi(\theta_1, \theta_2 \cdots \theta_B | y)$

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

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

proposal density : $q(\theta_i, \theta_i' | y, \theta_1, \underbrace{\theta_2 \cdots \theta_{i-1}}, \underbrace{\theta_{i+1} \cdots \theta_B})$

$$: q(\theta_i, \theta_i' | y, \psi_{i-1}, \psi^{i+1})$$

$\alpha(\theta_i, \theta_i' | y, \psi_{i-1}, \psi^{i+1})$ = probability of moving

$$= \min\left\{1, \frac{f(y | \theta_i', \psi_{i-1}, \psi^{i+1})\, \pi(\theta_i', \theta_{-i})\, q(\theta_i', \theta_i | y, \psi_{i-1}, \psi^{i+1})}{f(y | \theta_i, \psi_{i-1}, \psi^{i+1})\, \pi(\theta_i, \theta_{-i})\, q(\theta_i, \theta_i' | y, \psi_{i-1}, \psi^{i+1})}\right\}$$

We are concerned about $\pi(\theta_i^*, \theta_{-i}^* | y)$

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

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

for some i

$$\pi(\theta_i^* | y, \psi_{i-1}^*, \psi^{i+1})\, q(\theta_i^*, \theta_i | y, \psi_{i-1}^*, \psi^{i+1})\, \alpha(\theta_i^*, \theta_i | y, \psi_{i-1}^*, \psi^{i+1})$$

$$= \underbrace{\pi(\theta_i | y, \psi_{i-1}^*, \psi^{i+1})}_{\text{full conditional}}\, q(\theta_i, \theta_i^* | y, \psi_{i-1}^*, \psi^{i+1})\, \alpha(\theta_i, \theta_i^* | y, \psi_{i-1}^*, \psi^{i+1})$$

integrating and multiplying both sides by $\pi(\psi^{i+1} \mid y, \psi^*_{i-1})$

$$\iint\cdots\int \pi(\theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, q(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})\, \alpha(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})\, \pi(\psi^{i+1} \mid \psi^*_{i-1}, y)\, d\psi^{i+1}\, d\psi^*_{i-1}$$
$$= \iint\cdots\int \pi(\theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})\, q(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, \alpha(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, \pi(\psi^{i+1} \mid y, \psi^*_{i-1})\, d\psi^{i+1}\, d\psi^*_{i-1}$$

$$\iint\cdots\int \pi(\theta_i^* \mid y, \psi^*_{i-1})\, \pi(\psi^{i+1} \mid \psi_i^*, y)\, q(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})\, \alpha(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})\, d\psi^{i+1}\, d\psi^*_{i-1}$$
$$= \iiint \pi(\theta_i, \psi^{i+1} \mid \psi^*_{i-1}, y)\, q(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, \alpha(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, d\psi^{i+1}\, d\psi^*_{i-1}$$

$$\pi(\theta_i^* \mid y, \psi^*_{i-1}) = \frac{\iint\cdots\int \pi(\theta_i, \psi^{i+1} \mid \psi^*_{i-1}, y)\, q(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, \alpha(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, d\psi}{\iint\cdots\int \pi(\psi^{i+1} \mid \psi_i^*, y)\, q(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})\, \alpha(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})\, d\psi}$$

$$\pi(\theta_i^* \mid y) = \frac{E_1(\alpha(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, q(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})}{E_2(\alpha(\theta_i^*, \theta_i \mid y, \theta^*_{i-1}, \psi^{i+1})\, d\psi}$$

$$\iint\cdots\int \underbrace{\pi(\theta_i, \psi^{i+1} \mid \psi^*_{i-1}, y)}_{f(x)}\, \underbrace{q(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})\, \alpha(\theta_i, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1})}_{x}\, d\psi = E_1(x)$$

$$\iint\cdots\int \underbrace{\pi(\psi^{i+1} \mid \psi_i^*, y)\, q(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})}_{f(x)}\, \underbrace{\alpha(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1})}_{x}\, d\psi = E_2(x)$$

To estimate numerator and denominator, follow the following steps.

For numerator

1) Set $\psi_{i-1} = \psi^*_{i-1}$ and sample reduced set of full conditionals $\pi(\theta_k \mid y, \theta_{-k})$ for $k \geq i$ which is just drawing from MH algorithm as we don't directly know full conditionals.

$$\{\theta_i^{(g)} \dots \theta_B^{(g)}\}, \quad g = 1, \dots, M$$

For denominator

2) Let $\psi_i^* = (\theta_1^*, \theta_2^* \dots \theta_i^*)$. Again sample from full conditionals $\pi(\theta_l \mid y, \theta_{-l})$ for $l \geq i+1$ which is just drawing from MH algorithm.

$$\{\theta_{i+1}^{(j)} \dots \theta_B^{(j)}, \quad j = 1, \dots, J$$

Also draw $\theta_i^{(j)}$ from $q(\theta_i^*, \theta_i \mid y, \psi^*_{i-1}, \psi^{i+1\,(j)})$

3) Final estimate

$$\hat{\pi}(\theta_i^* \mid y, \theta_1^*, \dots \theta_{i-1}^*) =$$

$$\frac{\frac{1}{M}\sum_{g=1}^{M} \alpha(\theta_i^{(g)}, \theta_i^* \mid y, \psi_{i-1}^*, \psi^{i+1\,(g)})\, q(\theta_i^{(g)}, \theta_i^* \mid y, \psi^*_{i-1}, \psi^{i+1,(g)})}{\frac{1}{J}\sum_{j=1}^{J} \alpha(\theta_i^*, \theta_i^{(j)} \mid y, \psi_{i-1}^*, \psi^{i+1,(j)})}$$

Marginal likelihood estimator

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