Efficient MCMC Sampling Finite-State Markov Jump Processes and Bayesian Inference

Vinayak Rao*
Department of Statistics, Purdue University
and
Boqian Zhang
Department of Statistics, Purdue University

August 22, 2016

Abstract

Abstrct.

Keywords: Markov jump process, MCMC, Metropolis Hasting sampler, Bayesian inference

1 Introduction

In this paper, we tackled the problem of sampling MJP parameters from the posteriors, efficiently, using Metropolis Hasting algorithm.

2 Metropolis Hasting for Bayesian Inference using FFBS within the Gibbs Sampling On MJPs

*The authors gratefully acknowledge

3 Metropolis Hasting for Bayesian Inference using FFBS within the Gibbs Sampling On MJPs

4 Verifications of Algorithm 1

Proof of Algorithm 1:

Assume: $S = [S_0, S_1, ..., S_N]$, $T = [T_0, T_1, ..., T_N, T_{N+1}(T_{end})]$, and y as observations.

In JMLR-2013 Fast MCMC Sampling for MJP and Extensions, the FFBS frame contains α_t as follows.

Since after uniformization, the virtual jumps are added. Then the state process of the trajectory with virtual jumps is just a discrete time markov jump process. The key point is that we need to have W be conditioned, to get the marginal probability $P(y_{[T_0,T_{N+1})}|\theta, W)$ from FFBS algorithm.

$$\alpha_t^\theta(s) = P(S_t = s\ , y_{[T_0,T_t)}, U, T).$$

$$P(y_{[T_0,T_{N+1})}|\theta, W) = \sum_{s=0}^{N-1} \alpha_N^\theta(s) \cdot P(y_{[T_N,T_{N+1})}|S_N = s).$$

$$P(\theta, W|y) \propto P(\theta, W, y) = P(y|W, \theta)P(W|\theta)P(\theta).$$

$P(y|W, \theta)$ is the marginal probability we get after Forward Filtering Algorithm and the $P(W|\theta)$ is the probability density for the $poisson(\Omega)$, because of the uniformization procedure. Let denote the kernel for (a), (b) and (c) as $\kappa_1(\theta^*|\theta, W, T, S, y)$, $\kappa_2(S^*, T^*|S, T, W, \theta^*, y)$ and $\kappa_3(W^*|S^*, T^*, \theta^*, y)$. For Step (a) $\kappa_1(\theta^*|\theta, W, T, S)$:

$$\begin{aligned} P((W, T, S, \theta) \to (W, T, S, \theta^*))P(\theta, W|y) &= P(\theta^*, W|y)q(\theta|\theta^*) \wedge P(\theta, W|y)q(\theta^*|\theta) \\ &= P((W, T, S, \theta^*) \to (W, T, S, \theta))P(\theta^*, W|y). \end{aligned}$$

$\therefore \int \kappa_1(\theta^*|\theta)P(\theta, W|y)d\theta = P(\theta^*, W|y)$.

So the stationary distribution of κ_1 is $P(\theta, W|y)$.

Step (b) $\kappa_2(S^*, T^*|S, T, W, \theta^*, y)$:

Step(b) is the same as Fast MJPs Gibbs sampling scheme.

$$((S, T, \theta, W) \to (S^*, T^*, \theta, W)|y) = P(V^*|W, \theta, y) = P(V^*|W, \theta, y)/P(W, \theta, y)$$

$$\begin{aligned} P((S, T) \to (S^*, T^*)|W, \theta, y)P(S, T|W, \theta, y) &= P(V^*|W, \theta, y)P(V|W, \theta, y) \\ &= P((S^*, T^*) \to (S, T)|W, \theta, y)P(S^*, T^*|W, \theta, y) \end{aligned}$$

So the stationary distribution of $\kappa_2(S^*, T^*|S, T, W, y)$ is $P(S, T|W, \theta, y)$. Now, let's consider $\kappa_2 \circ \kappa_1(S^*, T^*, \theta^*|S, T, \theta, y, W)$.

$$((S, T, \theta, W) \to (S^*, T^*, \theta^*, W)|y) = P((W, T, S, \theta) \to (W, T, S, \theta^*))P((S, T, \theta^*.W) \to (S^*, T^*, \theta^*, W)|y).$$

The stationary distribution of $\kappa_1(S^*, T^*, U^*|S, T, U)$ is $P(S, T, U|\theta, y)$. And the stationary distribution of $\kappa_2(U^*|U)$ is $P(U|S, T, \theta, y)$.

$$\begin{aligned} &P((S, T, \theta, W) \to (S^*, T^*, \theta^*, W)|y)P(S, T, \theta|W, y) \\ &= P((W, T, S, \theta) \to (W, T, S, \theta^*)) \cdot P(\theta|W, y) \cdot P((S, T, \theta^*.W) \to (S^*, T^*, \theta^*, W)|y)P(S, T|\theta, W, y) \\ &= P((W, T, S, \theta^*) \to (W, T, S, \theta)) \cdot P(\theta^*|W, y) \cdot P((S^*, T^*, \theta^*.W) \to (S, T, \theta^*, W)|y)P(S^*, T^*|\theta, W, y) \\ &= P((S^*, T^*, \theta^*, W) \to (S, T, \theta, W)|y)P(S, T, \theta|W, y). \end{aligned}$$

So the stationary distribution of $\kappa_2 \circ \kappa_1$ is $P(S, T, \theta|W, y)$.

Obviously, $\kappa_3(W^*|W, S^*, T^*, \theta^*, y)$ has $P(W|S^*, T^*, \theta^*, y)$ as stationary distribution.

Therefore, $\int \kappa_3(W^*|W, S^*, T^*, \theta^*, y)P(W, S^*, T^*, \theta^*|y)dW = P(W^*, S^*, T^*, \theta^*|y)$.

Thus, $\int \kappa_3 \cdot (\int \kappa_2 \circ \kappa_1 \cdot P(W, S, T, \theta|y)d\theta dS dT)dW = \int \kappa_3 P(W, S^*, T^*, \theta^*|y)dW = P(W^*, S^*, T^*, \theta^*|y)$.

So the stationary distribution of $\kappa_3 \circ \kappa_2 \circ \kappa_1$ is $P(W^*, S^*, T^*, \theta^*|y)$.

5 Verifications of Algorithm 2

Proof: Our state is $(W, S, T, \theta, \theta^*)$.

$$
\begin{aligned}
p(y, W, S, T, \theta, \theta^*) &= p(\theta)q(\theta^*|\theta)P(S, T|\theta, \theta^*)P(W|S, T, \theta, \theta^*)P(y|S, T, \theta, \theta^*) \\
&= p(\theta)q(\theta^*|\theta)P(S, T|\theta)P(W|S, T, \theta, \theta^*)P(y|S, T).
\end{aligned}
$$

The marginal distribution of $(y, S, T, \theta, \theta^*)$ and (y, S, T, θ) as follows.

$$
\begin{aligned}
p(y, S, T, \theta, \theta^*) &= p(\theta)q(\theta^*|\theta)P(S, T|\theta, \theta^*)P(y|S, T, \theta, \theta^*) \\
&= P(y, S, T, \theta)q(\theta^*|\theta).
\end{aligned}
$$

$$
p(y, S, T, \theta) = p(\theta)P(S, T|\theta)P(y|S, T, \theta).
$$

So the conditional distribution over θ^* given (y, S, T, θ) is $q(\theta^*|\theta)$. And the conditional distribution over W given $(y, S, T, \theta, \theta^*)$ is $P(W|S, T, \theta, \theta^*)$, which is actually the distribution of Non Homogeneous Poisson Process with piecewise constant rate $h(\theta) + h(\theta^*) - A_{S(t)}(\theta)$.
Thus the Step 1 + Step 2 is actually equivalent to sampling from the conditional distribution $P(\theta^*, W|S, T, \theta, y)$.
The Step 3 + Step 4 satisfy the detailed balance condition. The reason is as follows.

$$
\begin{aligned}
&P((W, S, T, (\theta, \theta^*)) \to (W, S^*, T^*, (\theta^*, \theta)))P(S, T, (\theta, \theta^*)|W, y) \\
&= (1 \wedge \frac{P((\theta^*, \theta)|W, y)}{P((\theta, \theta^*)|W, y)})P(S^*, T^*|W, (\theta^*, \theta), y)P(S, T|W, (\theta, \theta^*), y)P((\theta, \theta^*)|W, y) \\
&= P((W, S^*, T^*, (\theta^*, \theta)) \to (W, S, T, (\theta, \theta^*)))P(S^*, T^*, (\theta^*, \theta)|W, y)
\end{aligned}
$$

Therefore the stationary distribution of this MCMC sampler is $P(W, S, T, (\theta, \theta^*)|y)$. Thus the stationary distribution of (S, T, θ) is the corresponding marginal distribution $P(S, T, \theta|y)$.

6 Exponential Model which do not have conjugate posterior

Assume: $S = [S_0, S_1, ..., S_N]$,$T = [t_0(t_{start}), t_1, ..., t_N, t_{N+1}(t_{end})]$, and y as observations.

We consider a specific structure of rate matrix A. $A_{ij} = \alpha f_{ij}(\beta),\ i \neq j.$ $A_{ii} = -\sum_{j\neq i} A_{ij}$. $0 \leqslant f_{ij} \leqslant 1$. Denote $F_i(\beta) = \sum_{j\neq i} f_{ij}(\beta)$.

$$
\begin{aligned}
P(s_0, S, T|\alpha, \beta) &= \pi_0(s_0) \prod_{i=1}^{N} A_{S_{i-1}S_i} \exp(-\int_{t_{start}}^{t_{end}} |A_{S(t)}|dt) \\
&= \pi_0(s_0)\alpha^N \prod_{i=1}^{N} f_{S_{i-1}S_i} \exp(-\alpha \sum_{i=0}^{N} F_{S_i}(\beta)(t_{i+1} - t_i))
\end{aligned}
$$

Assume the prior distributions of α, β are $p_1(\alpha)$ and $p_2(\beta)$.

Then the posterior distribution of parameters α, β will be as follows.

$$
P(\alpha, \beta|s_0, S, T) \propto \alpha^N \prod_{i=1}^{N} f_{S_{i-1}S_i} \exp(-\alpha \sum_{i=0}^{N} F_{S_i}(\beta)(t_{i+1} - t_i)) p_1(\alpha) p_2(\beta)
$$

If we assume the priors of α, β are $Gamma(\mu, \lambda)$, $Gamma(\omega, \theta)$, then the posterior will have a simper form as follows.

$$
P(\alpha, \beta|s_0, S, T) = C\alpha^{\mu+N-1} \exp(-\alpha(\lambda + \sum_{i=0}^{N} F_{S_i}(\beta)(t_{i+1} - t_i))) \prod_{i=1}^{N} f_{S_{i-1}S_i} \beta^{\omega-1} \exp(-\theta\beta)
$$

We notice that given $\beta,\ S,\ T,\ \alpha$ is distributed as a $Gamma$ distribution.

$\alpha|\beta, S, T, y = \alpha|\beta, S, T \sim Gamma(\mu + N, \lambda + \sum_0^N F_{S_i}(\beta)(t_{i+1} - t_i))$.

But there is no conjugate distribution to sample $\beta \sim P(\beta|s_0, S, T)$. We will have to use Metropolis Hasting within Gibbs to sample β.

$$
P(\beta|S, T) = C \frac{\prod_{i=1}^{N} f_{S_{i-1}S_i}(\beta) \beta^{\omega-1} \exp(-\theta\beta)}{(\lambda + \sum_{i=0}^{N} F_{S_i}(\beta)(t_{i+1} - t_i))^{\mu+N}}
$$

7 Immigration models with capability

Now, let's consider a immigration model as follows. We have state space$0, 1, 2, ..., N$, representing the total population. The transition matrix is defined as follows.

$$A_i =: A_{i,i} = -(\alpha + i\beta), \quad i = 0, 1, ..., N$$

$$A_{i,i+1} = \alpha, \quad i = 0, 1, ..., N-1,$$

$$A_{i,i-1} = \beta, \quad i = 1, ..., N.$$

We already know the conditional density(given α, β) of a MJP trajectory (s_0, S, T) in time interval $[t_{start}, t_{end}]$, with $S = (s_1, s_2, ..., s_k)$, $T = (t_1, t_2, ..., t_k)$.

$$f(s_0, S, T|\alpha, \beta) = \prod_{i=0}^{k-1} A_{s_i,s_{i+1}} \exp(\sum_{i=0}^{k} A_{s_i}(t_{i+1} - t_i)),$$

where $t_0 = t_{start}$, $t_{k+1} = t_{end}$.

Let's denote some notations here.

$$U(s_0, S, T) := \sum_{i=0}^{k-1} \mathbb{I}_{\{s_{i+1}-s_i=1\}}$$

$$D(s_0, S, T) := \sum_{i=0}^{k-1} \mathbb{I}_{\{s_{i+1}-s_i=-1\}}$$

Call them U and D for short. Let's denote the total time when the trajectory state stays at state i as τ_i, i.e. $\tau_i = \sum_{j=0}^{k}(t_{j+1} - t_j)\mathbb{I}_{\{s_j=i\}}$, then $\sum_{i=0}^{k}(t_{i+1} - t_i)s_i = \sum_{i=0}^{N} \tau_i i$

$$f(s_0, S, T|\alpha, \beta) = \exp(-\alpha(t_{end} - t_{start} - \tau_N))\alpha^U \cdot \exp((-(\sum_{i=0}^{k}(t_{i+1} - t_i)s_i)\beta) \prod_{i=1}^{N} i^{\sum_{j=0}^{k-1} \mathbb{I}_{s_{j+1}=i-1}, {}_{s_j=i}} \beta^D$$

If we assume the prior of α, and β are $Gamma(\mu, \lambda)$, $Gamma(\omega, \theta)$, which are independent with each other.

$$p(\alpha) = \frac{\lambda^\mu}{\Gamma(\mu)} \alpha^{\mu-1} e^{-\lambda\alpha}$$

.

$$p(\beta) = \frac{\theta^{\omega}}{\Gamma(\omega)}\beta^{\omega-1}e^{-\theta\beta}$$

. Then we can get the posterior distribution

$$f(\alpha, \beta|s_0, S, T)$$

as follows.

$$f(\alpha, \beta|s_0, S, T) \propto \exp(-(\lambda + t_{end} - t_{start} - \tau_N)\alpha)\alpha^{\mu+U-1} \cdot \exp(-(\sum_{i=0}^{k}(t_{i+1} - t_i)s_i + \theta)\beta)\beta^{\omega+D-1}.$$

It means that the posterior distributions of α, β are still independent.
$\alpha|s_0, S, T$ is following $Gamma(\mu + U, \lambda + t_{end} - t_{start} - \tau_N)$
$\beta|s_0, S, T$ is following $Gamma(\omega + D, \theta + \sum_{i=0}^{k}(t_{i+1} - t_i)s_i)$, which is equivalent to $Gamma(\omega + D, \theta + \sum_{i=0}^{N} \tau_i i)$

8 Conclusion

SUPPLEMENTAL MATERIALS

Title: Brief description. (file type)

R-package for MYNEW routine: R-package MYNEW containing code to perform the diagnostic methods described in the article. The package also contains all datasets used as examples in the article. (GNU zipped tar file)

HIV data set: Data set used in the illustration of MYNEW method in Section 3.2. (.txt file)

References

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics* **32**, 159–188.

Algorithm 1 MH In Gibbs sampling for MJPs

Input: A set of partial and noisy observations $y_{[t_0,t_{N+1})}$, Initial distribution over states π_0, Metropolis Hasting proposal $q(.|\theta)$.

The previous MJP path $S(t) = (S, T)$, the previous MJP parameters θ.

Output: A new MJP trajectorie $\tilde{S}(t) = (\tilde{S}, \tilde{T})$, A series of MJP parameters $\tilde{\theta}$.

0: Let $\Omega = h(\theta)$, with $\Omega > max_s|A_s|$ using some deterministic function h.

1: Sample virtual jumps $U \subset [t_{start}, t_{end}]$ from a Non homogeneous Poisson process with piecewise-constant rate

$$R(t) = (\Omega + A_{S(t)}).$$

Define $W = T \cup U$.

2: Propose $\theta^* \sim q(.|\theta)$.

Accept θ^* as $\tilde{\theta}$ with probability α.

$$\begin{aligned}\alpha &= 1 \wedge \frac{P(W, \theta^*|y)}{P(W, \theta|y)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \\ &= 1 \wedge \frac{P(y|W, \theta^*)P(W|\theta^*)p(\theta^*)}{P(y|W, \theta)P(W|\theta)p(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}.\end{aligned}$$

3: Sample a path $\tilde{V}$, from a discret-time Markov chain with $|W|+1$ steps, using FFBS algorithm. The transition matrix of the Markov chain is $B = (I + \frac{A}{\Omega})$ while the initial distribution over states is π_0. The likelihood of state s at step i is

$$L_i(s) = P(Y_{[w_i,w_{i+1})}|S(t) = s \; for \; t \in [w_i, w_{i+1})) = \prod_{j:t_j \in [w_i,w_{i+1})} p(y_{t_j}|S(t_j) = s).$$

4: Let $\tilde{T}$ be the set of times in W when the Markov chain changes state. Define $\tilde{S}$ as the corresponding set of state values. Return $(\tilde{S}, \tilde{T}, \tilde{\theta})$.

Algorithm 2 MH In Gibbs sampling for MJPs

Input: A set of partial and noisy observations $y_{[t_0,t_{N+1})}$, Initial distribution over states π_0, Metropolis Hasting proposal $q(.|\theta)$.

The previous MJP path $S(t) = (S, T)$, the previous MJP parameters (θ).

Output: A new MJP trajectorie $\tilde{S}(t) = (\tilde{S}, \tilde{T})$, A series of MJP parameters $\tilde{\theta}$.

0: Sample $\theta^* \sim q(.|\theta)$. And let $\Omega = h(\theta)+h(\theta^*)$, with $h(\theta) > max_s|A_s(\theta)|$, $h(\theta^*) > max_s|A_s(\theta^*)|$ using some deterministic function h.

1: Sample virtual jumps $U \subset [t_{start}, t_{end}]$ from a Non homogeneous Poisson process with piecewise-constant rate

$$R(t) = (\Omega + A_{S(t)}(\theta)).$$

Define $W = T \cup U$.

2: Propose (θ^*, θ) and accept θ^* as $\tilde{\theta}$ with probability α.

$$\begin{aligned}
\alpha &= 1 \wedge \frac{P(W, (\theta^*, \theta)|y)}{P(W, (\theta, \theta^*)|y)} \\
&= 1 \wedge \frac{P(y|W, \theta^*, \theta)P(W|(\theta^*, \theta))p((\theta^*, \theta))}{P(y|W, (\theta, \theta^*))P(W|(\theta, \theta^*))p((\theta, \theta^*))} \\
&= 1 \wedge \frac{P(y|W, \theta^*, \theta)p((\theta^*, \theta))}{P(y|W, (\theta, \theta^*))p((\theta, \theta^*))}.
\end{aligned}$$

3: Sample a path $\tilde{V}$, from a discret-time Markov chain with $|W|+1$ steps, using FFBS algorithm. The transition matrix of the Markov chain is $B = (I + \frac{A(\tilde{\theta})}{\Omega})$ while the initial distribution over states is π_0. The likelihood of state s at step i is

$$L_i(s) = P(Y_{[w_i,w_{i+1})}|S(t) = s \; for \; t \in [w_i, w_{i+1})) = \prod_{j:t_j \in [w_i,w_{i+1})} p(y_{t_j}|S(t_j) = s).$$

4: Let $\tilde{T}$ be the set of times in W when the Markov chain changes state. Define $\tilde{S}$ as the corresponding set of state values. Return $(\tilde{S}, \tilde{T}, \tilde{\theta})$.

Algorithm 3 Generic Gibbs sampling for MJPs for Gamma priors

Input: observations $y_{[t_0,t_{k+1})}$

Initialize, $i = 0$

(a) Set $\alpha(0), \beta(0)$ arbitrarily and set current trajectory $[S,T](0)$ arbitrarily.

(b) Uniformize $[S,T](0)$, to get virtual jumps U.

repeat

for $i = 1$ **to** N **do**

(a) Sample $U(i) \sim P(U|\beta(i-1), \alpha(i-1), S(i-1), T(i-1), y)$.

(b) Use FFBS algorithm to sample states given all the jump times(both true jumps and virtual jumps). (i.e. $V(i) \sim P(V|\beta(i-1), \alpha(i-1), W(i), y)$.) Then delete all the virtual jumps to get $S(i), T(i)$.

(c) Propose $\beta^* \sim q(.|\beta(i-1))$.

Set $\beta(i) = \beta^*$, with probability $P_{acc} = 1 \wedge \frac{P(\beta^*|S(i),T(i))}{P(\beta(i-1)|S(i),T(i))} \frac{q(\beta(i-1)|\beta^*)}{q(\beta^*|\beta(i-1))}$;

Otherwise set $\beta(i) = \beta(i-1)$.

(d) Sample $\alpha(i) \sim P(.|\beta(i), S(i), T(i), y)$.

It is a $Gamma(\mu + N, \lambda + \sum_0^N F_{S_i}(\beta)(t_{i+1} - t_i))$ distribution actually.

end for

until $i = N$