BOQIAN MJPS*

AUTHOR†

Abstract.

Key words. MCMC, Markov Jump Process, Particle MCMC, Slice Sampling, infinite state space

1. Details of Fast Gibbs MJPs Iteration.

Proof. Denote the Virtual jump time as U and, true jumps as S, and true jump time as T. Denote the transition kernel of step 2 and step 3 as $\kappa_2(S^*, T^*|S, T, W, y)$.

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

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

So the stationary distribution of $\kappa_2(S^*, T^*|S, T, W, y)$ is $P(S, T|W, y)$. Then combine it with the transition kernel of step $1\kappa_1(W^*|S, T, y)$. So the stationary distribution of $\kappa_2 \circ \kappa_1$ is $P(S, T, W|y)$.

$\square$

2. Details of FFBS in Gibbs Sampling on MJPs.

Suppose the state space is finite and has N states, i.e. $\mathbb{S} = \{0, 1, 2, ..., N-1\}$. And the trajectory is $S = [s_1, s_2, ..., s_K]$, and $T = [t_0, t_1, ..., t_K, t_{K+1}]$, where $t_0 = t_{start}$, $t_{K+1} = t_{end}$. Beside, we observe $y = [y_1, y_2, ..., y_o]$, at time $\tau = [\tau_1, \tau_2, ..., \tau_o]$. We are interested in sampling from posterior distribution $P(s_0, S|T, y_{[0,t_{end}]})$. The posterior distribution of s_0, S is just a markov jump process, with transition matrix $B = \frac{A}{\Omega} + I$.

Forward Filtering:

Define $\alpha_i(s) = P(S_i = s, y_{[t_0,t_i]}) = \sum_{v=0}^{N} \alpha_{i-1}(v)P(y_{[t_{i-1},t_i]}|S_{i-1} = v)P(S_i = s|S_{i-1} = v) = \sum_{v=0}^{N} \alpha_{i-1}(v)L_{i-1}(v)P(S_i = s|S_{i-1} = v)$.

$\alpha_0(s) = \mu(s)$, which is the initial distribution.

$L_i(s) = P(y_{[t_i,t_{i+1}]}|S_i = s) = \prod_{i=1}^{o} f(y_i|s)^{\mathbb{I}\{\tau_i \in [t_i, t_{i+1})\}}$, $i = 0, 1, 2, ..., K$ where $f(.|.)$ is the observation probability density. Backward Sampling:

Define $\beta_i(s) = P(S_i = s|S_{i+1}, y[0, t_{end}))$, $i = 0, 1, ..., K-1$.

$\beta_K(s) = P(S_K = s|y[0, t_{end}))$.

$$\beta_K(s) \propto \alpha_K(s) \cdot L_K(s).$$

$$\beta_i(s) \propto \alpha_i(s) \cdot L_i(s) \cdot B_{sS_{i+1}}.$$

Sample the new states backwardly.

3. Details of FFBS in Continuous time Gibbs Sampling on MJPs.

We use the same notations as the previous section.

Since in MCMC for continuous-time discrete-state systems[2012 NIPS], there is no uniformization procedure, so the posterior $P(s_0, S|T, y_{[0,t_{end}]})$ is no longer a discrete-time markov jump process.

*This work was supported by the Society for Industrial and Applied Mathematics

†Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania. (tex@siam.org). Questions, comments, or corrections to this document may be directed to that email address.

Denote $\Delta t_i = ti+1-t_i$.
Now, we define $\alpha_i(s) = P(S_i = s, t_{0:i}, y_{[t_0,t_i]}) = \sum_{v=0}^{N} \alpha_{i-1}(v) P(y_{[t_{i-1},t_i]}|S_{i-1} = v)P(\Delta t_{i-1}|S_{i-1} = v)P(S_i = s|S_{i-1} = v) = \sum_{v=0}^{N} \alpha_{i-1}(v)L_{i-1}(v)P(S_i = s|S_{i-1} = v)$.
Now, $L_i(s) = P(y_{[t_i,t_{i+1}]}|S_i = s)P(\Delta t_i|S_i = s) = \prod_{i=1}^{o} f(y_i|s)^{\mathbb{I}\{\tau_i \in [t_i,t_{i+1})\}} P(\Delta t_i|S_i = s),\ i = 0,1,2,...,K-1$ where $f(.|.)$ is the observation probability density.
$L_K(s) = P(y_{[t_K,t_{K+1}]}|S_i = s)e^(B_{S_K}\Delta t_k)$.
Let $C = I - Diag(B_0, B_1, ..., B_{N-1})^{-1}A$.

$$\beta_K(s) \propto \alpha_K(s)L_K(s).$$

$$\beta_i(s) \propto \alpha_i(s) \cdot L_i(s) \cdot C_{s,S_{i+1}}.$$

Sample the new states backwardly.

4. SMC for MJPs.

Algorithm:
Basic Assumptions: We are considering the time interval $[0,T]$, and we have N observations $y_1, y_2, ..., y_N$, at time $t_1, t_2, ..., t_N$. We assume observing time $t_1, t_2, ..., t_N$ are known at first. The parameter$\theta = (A, u(.), P(Y_n = .|x_n = x))$. A is the transition matrix of the jump process.
Let's denote our proposal distribution as $q_\theta(.|y_{1:N})$.
Step 1.
At time $n = 1$.
Sample N particles $X^k_{[0,t_1]}, k = 1,2,...,N$.
Compute and normalize the weights.

$$w_1(X^k_{[0,t_1]}) := \frac{P_\theta(X^k_{[0,t_1]}, y_1)}{q_\theta(X^k_{[0,t_1]}|y_1)} = \frac{\mu_\theta(X^k_{[0,t_1]})g_\theta(y_1|X^k_{[0,t_1]})}{q_\theta(X^k_{[0,t_1]}|y_1)}$$

$$W_1(X^k_{[0,t_1]}) = \frac{w_1(X^k_{[0,t_1]})}{\sum_{i=1}^{N} w_1(X^i_{[0,t_1]})}$$

Step 2.
At time $n = 2,3,4,...,P$
(a) Sample $A^k_{n-1} \sim Multi(.|W_{n-1})$.
(b) Sample $X^k_{[t_{n-1},t_n]} \sim q^n_\theta(.|y_n, X^{A^k_{n-1}}_{[0,t_{n-1}]})$.
Then, set $X^k_{[0,t_n]} := combine(X^{A^k_{n-1}}_{[0,t_{n-1}]}, X^k_{[t_{n-1},t_n]})$.
(c) Calculate the weights and normalize the weights.

$$w_n(X^k_{[0,t_n]}) := \frac{P_\theta(X^k_{[0,t_n]}, y_{1:n})}{P_\theta(X^{A^k_{n-1}}_{[0,t_{n-1}]}, y_{1:n})q^n_\theta(X^k_{[t_{n-1},t_n]}|y_{1:n}, X^{A^k_{n-1}}_{[0,t_{n-1}]})} = \frac{f_\theta(X^k_{[t_{n-1},t_n]}|X^{A^k_{n-1}}_{[0,t_{n-1}]})g(y_n|X^k_{t_n})}{q^n_\theta(X^k_{[t_{n-1},t_n]}|y_{1:n}, X^{A^k_{n-1}}_{[0,t_{n-1}]})}$$

$$W_n(X^k_{[0,t_n]}) = \frac{w_n(X^k_{[0,t_n]})}{\sum_{i=1}^{N} w_n(X^i_{[0,t_n]})}$$

Specify $f_\theta, g_\theta, q_\theta^n$:
1. Let $X_{[t,t']}$ be the trajectory that is equivalent to $S = s_0, s_1, ..., s_n$, $T = t_1, t_2, ..., t_n$.
2. $f_\theta(X_{[t,t']}|X_{[0,t]}) = f_\theta(X_{[t,t']}|X_t = s_0^*) = A_{s_0^* s_1} e^{-|A_{s_0^*}|(t_1-t)} \prod_{i=2}^n (|A_{s_{i-1}}|e^{-|A_{s_{i-1}}|(t_i-t_{i-1})} A_{s_{i-1}s_i}) e^{|A_{s_n}|(t'-t_n)}$.
3.$g_\theta(y_n|X_{[t',t_n]}) = G(y_n|X_{t_n})$.
4.q_θ^n is the important sampling proposal distribution. In our case, we just let it be the following.
$q_\theta^n(.|y_n, X_{[0,t_{n-1}]}) = f_\theta(X_{[t_{n-1},t_n]}|X_{t_{n-1}})$.
5. $\mu(X_{[0,t]}) = \pi_0(s_0) A_{s_0^* s_1} \prod_{i=2}^n (|A_{s_{i-1}}|e^{-|A_{s_{i-1}}|(t_i-t_{i-1})} A_{s_{i-1}s_i}) e^{|A_{s_n}|(t-t_n)}$, with $t_0 = 0$

This procedure provides us at time T with an approximation of the joint posterior density $p_\theta(dX_{[0,T]}|y_{1:n})$ given by

$$\hat{p_\theta}(dX_{[0,T]}|y_{1:n}) = \sum_{k=1}^N W_n^k \delta_{X^k[0:T]}(dX_{[0,T]})$$

In addition, the estimate of the marginal likelihood $p_\theta(y_{1:n})$ is given by

$$\hat{p}_\theta(y_{1:n}) = \hat{p}_\theta(y_1) \prod_{i=2}^n \hat{p}_\theta(y_i|y_{i-1})$$

where

$$\hat{p}_\theta(y_i|y_{i-1}) = \frac{1}{N} \sum_{k=1}^N w_n(X_{[0,t_i]})$$

is an estimate computed at time i of

$$p_\theta(y_i|y_{i-1}) = \int w_n(X_{[0,t_i]}) q_\theta(X_{[t_{i-1},t_i]}|y_i, X_{[0,t_{i-1}]}) p_\theta(X_{[0,t_{i-1}]}|y_{1:i-1}) dX_{[0,t_i]}.$$

5. PMCMC On MJPs.

Algorithm:
Basic Assumptions: Here, we have N observations $y_1, y_2, ..., y_N$, at time $t_1, t_2, ..., t_N$. The time interval we are interested in is $[0, T]$. The model is Hidden Markov State model.
Targeting distribution: $P(X_{[0:T]}|y_{1:N})$.
Step1:
Initialize $i = 0$. Run a SMC algorithm targeting $P(X_{[0:T]}|y_{1:N})$. Sample a trajectory $X_{[0:T]}(0) \sim \hat{P}_\theta(.|y_{1:N})$. We let $\hat{P}_\theta(y_{1:N})(0)$ denote the corresponding marginal likelihood estimate.
Step2:
For iteration $i \geqslant 1$,
(a). Run a SMC targeting sample $X_{[0:T]}^* \sim \hat{P}_\theta(.|y_{1:N})$, and let $\hat{P}_\theta(y_{1:N})^*$ denote the marginal likelihood estimate.
(b). With probability $1 \wedge \frac{\hat{P}_\theta(y_{1:N})^*}{\hat{P}_\theta(y_{1:N})(i-1)}$ to update $X_{[0:T]}(i) = X_{[0:T]}^*$, $\hat{P}_\theta(y_{1:N})(i) = \hat{P}_\theta(y_{1:N})^*$. Otherwise , $X_{[0:T]}(i) = X_{[0:T]}(i-1)$, $\hat{P}_\theta(y_{1:N})(i) = \hat{P}_\theta(y_{1:N})(i-1)$.

6. 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 $Gamma(\omega+D,\theta+\sum_{i=0}^{N}\tau_i i)$

7. Metropolis Hasting using FFBS within the Gibbs Sampling On MJPs(marginal mcmc version).
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] FFBS frame, we have defined 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 $U(\textit{virtual jump times})$ and $T(\textit{true jump times})$ be conditioned, to get the marginal probability $P(y_{[T_0,T_{N+1})}|\theta,U,T)$ from FFBS algorithm.

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

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

Algorithm:
Step1: Initialization, $i=0$,

(a) Set $\theta(0)$ arbitrarily and set current trajectory $[S,T](0)$ arbitrarily;
(b) Uniformize $[S,T](0)$, get a uniformized trajectory $[V,W](0)$
(c) Run FFBS algorithm to get the marginal probability $P(y_{[T_0,T_{N+1})}|\theta(0))$.

Step2: for iteration $i\geqslant 1$,

(a) Propose $\theta^*\sim q(.|\theta(i-1))$
(b) Sample Virtual jump times $U^*\sim P(.|\theta(i-1),[S,T](i-1),y_{[T_0,T_{N+1}]})$
(c) Sample W^* $P(.|\theta^*,U^*,T(i-1),y_{[T_0,T_{N+1})})$, with FFBS algorithm, and recording the marginal probability $P(y_{[T_0,T_{N+1})}|\theta^*)$. $[S^*,T^*]$ is the new proposed trajectory.
(d) With probability

$$1\wedge\frac{P(y_{[T_0,T_{N+1})}|\theta^*)p(\theta^*)}{P(y_{[T_0,T_{N+1})}|\theta(i-1))p(\theta(i-1))}\frac{q(\theta(i-1)|\theta^*)}{q(\theta^*|\theta(i-1))}$$

set $\theta(i)=\theta^*$, $[S,T](i)=[S^*,T^*]$, and $P(y_{[T_0,T_{N+1})}|\theta(i))=P(y_{[T_0,T_{N+1})}|\theta^*)$; Otherwise, set $\theta(i)=\theta(i-1)$, $[S,T](i)=[S,T](i-1)$, and $P(y_{[T_0,T_{N+1})}|\theta(i))=P(y_{[T_0,T_{N+1})}|\theta(i-1))$.

8. Proposed Metropolis Hasting within the Gibbs Sampling On MJPs.

Denote all the parameters as θ, observations as Y.
Algorithm(iteration part):

Step1: Sample $U^* \sim P(.|\theta, S, T, Y)$.
Step2: Propose $\theta^* \sim q(.|\theta)$. , with acceptance rate α_0

$$\alpha_0 = \frac{P(\theta^*, S, U^*, T)}{P(\theta, S, U^*, T)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \wedge 1$$

Step3: Sample $W^* \sim P(.|\theta, U^*, T, Y)$, using FFBS.

Details about acceptance rate:
We are considering a Immigration model (α, β) with capacity N.
Assume: $S = [s_0, s_1, ..., s_N]$, $T = [t_0, t_1, ..., t_N, t_{N+1}(t_{end})]$.
Also assume there are u_i virtual jumps in time interval $[t_i, t_i + 1)$, $i = 0, 1, 2, ..., N$.

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

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

Call them U and D for short.
The proposal density $q(\theta^*|\theta) = \frac{1}{\sqrt{2\pi}\sigma\alpha^*} \exp(-(\frac{log(\alpha^* - \alpha)}{2\sigma^2})^2) \cdot \frac{1}{\sqrt{2\pi}\sigma\beta^*} \exp(-(\frac{log(\beta^* - \beta)}{2\sigma^2})^2)$

$$P(\theta, S, U, T) = \pi_0(s_0) \prod_{i=1}^{N} A_{s_{i-1}, s_i} \exp((t_0 - t_{N+1})\Omega) \prod_{i=0}^{N} (\Omega + A_{s_i})^{u_i}$$

$$\frac{P(\theta^*, S, U, T)}{P(\theta, S, U, T)} = \exp((t_0 - t_{N+1})(\Omega^* - \Omega)) \cdot (\frac{\alpha^*}{\alpha})^U (\frac{\beta^*}{\beta})^D \cdot \prod_{i=0}^{N} (\frac{\Omega^* + A^*_{S_i}}{\Omega + A_{S_i}})^{u_i}$$

$$\alpha_0 = \exp((t_0 - t_{N+1})(\Omega^* - \Omega)) \cdot (\frac{\alpha^*}{\alpha})^U (\frac{\beta^*}{\beta})^D \cdot \prod_{i=0}^{N} (\frac{\Omega^* + A^*_{S_i}}{\Omega + A_{S_i}})^{u_i} \cdot \frac{\alpha^* \beta^*}{\alpha\beta} \wedge 1$$

9. Metropolis Hasting using FFBS within the Gibbs Sampling On MJPs.

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

Proof.

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

Algorithm 1 MH In Gibbs sampling for MJPs

Input: observations $y_{[t_0,t_{k+1})}$
Initialize, $i = 0$
(a) Set $\theta(0)$ 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) Propose $\theta^* \sim q(.|\theta(i-1))$.
    Set $\theta(i) = \theta^*$, with probability $\alpha = 1 \wedge \frac{P(W,\theta^*|y)}{P(W,\theta(i-1)|y)} \frac{q(\theta(i-1)|\theta^*)}{q(\theta^*|\theta(i-1))}$;
    Otherwise set $\theta(i) = \theta(i-1)$.
    (b) Use FFBS algorithm to sample states given all the jump times(both true jumps and virtual jumps). (i.e. $V(i) \sim P(V|\theta(i), W(i-1), y)$.) Then delete all the virtual jumps to get $S(i), T(i)$.
    (c) Sample $U(i) \sim P(U|\theta(i), S(i), T(i), y)$.
  end for
until $i = N$

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.
So $\int \kappa_3(W^*|W, S^*, T^*, \theta^*, y)P(W, S^*, T^*, \theta^*|y)dW = P(W^*, S^*, T^*, \theta^*|y)$.
So $\int \kappa_3 \cdot (\int \kappa_2 \circ \kappa_1 \cdot P(W, S, T, \theta|y)d\theta dS dT)dW = \int \kappa_3 P(W, S^*, T^*, \theta^*|y)dW = P(W^*, S^*, T^*, \theta^*|y)$.
So the stationary distribution of $\kappa_3 \circ \kappa_2 \circ \kappa_1$ is $P(W^*, S^*, T^*, \theta^*|y)$.
$\square$

10. Hamiltonian MCMC On immigration model MJPs.

As we can see from last section, we can use Metropolis Hasting algorithm to sample $\theta \sim P(\theta|W, y)$. Hamiltonian MCMC can be applied to improve the acceptance rate. With HMC, a state proposed in this way can be distant from the current state but nevertheless have a high probability of acceptance.
Denote $\theta = (\alpha, \beta)^\top$ and $\alpha_t^\theta(s) = P(S_t = s, y_{[T_0,T_t)}|W)$, and $L_t(s) = P(y_{[T_t,T_{t+1})}|S_t = s)$ and the transition matrix of V as $B = I + \frac{A}{\Omega}$, with $\Omega = max\{\beta\ , \alpha\} + (d-2)\beta$. And $L_t(s)$ has nothing to do with θ
Now, consider the derivative $\nabla_\theta P(y|W, \theta)$.
We already know $P(y_{[T_0,T_{N+1})}|W, \theta) = \sum_{s=0}^{d-1} \alpha_N^\theta(s)L_N(s)$.
Since

$$\begin{aligned}
\alpha_t^\theta(s) &= P(S_t = s, y_{[T_0,T_t)}|W). \\
&= \sum_{v=0}^{d-1} \alpha_{t-1}^\theta(v)L_{i-1}(v)P(S_i = s|S_{i-1} = v). \\
&= \sum_{v=0}^{d-1} \alpha_{t-1}^\theta(v)L_{i-1}(v)B_{vs}.
\end{aligned}$$

So we can get the following updating equations.

$$\nabla_\theta P(y_{[T_0,T_{N+1})}|W,\theta) = \sum_{s=0}^{d-1} \nabla_\theta \alpha_N^\theta(s) L_N(s). \tag{10.1}$$

$$\nabla_\theta \alpha_t^\theta(s) = \sum_{v=0}^{d-1} (\nabla_\theta \alpha_{t-1}^\theta(v) + \alpha_{t-1}^\theta(v)\nabla_\theta B_{vs}) L_{t-1}(v). \tag{10.2}$$

$$\nabla_\theta \alpha_0^\theta(s) = (0,0)^\top. \tag{10.3}$$

(10.4)

So if we consider $\frac{\partial}{\partial\alpha}$, $\frac{\partial}{\partial\beta}$ separately, then we can get the following updating equations.

$$\frac{\partial}{\partial\alpha} P(y_{[T_0,T_{N+1})}|W,\theta) = \sum_{s=0}^{d-1} \frac{\partial}{\partial\alpha} \alpha_N^\theta(s) L_N(s). \tag{10.5}$$

$$\frac{\partial}{\partial\beta} P(y_{[T_0,T_{N+1})}|W,\theta) = \sum_{s=0}^{d-1} \frac{\partial}{\partial\beta} \alpha_N^\theta(s) L_N(s). \tag{10.6}$$

$$\frac{\partial}{\partial\alpha} \alpha_t^\theta(s) = \sum_{v=0}^{d-1} (\frac{\partial}{\partial\alpha} \alpha_{t-1}^\theta(v) + \alpha_{t-1}^\theta(v)\frac{\partial}{\partial\alpha} B_{vs}) L_{t-1}(v). \tag{10.7}$$

$$\frac{\partial}{\partial\beta} \alpha_t^\theta(s) = \sum_{v=0}^{d-1} (\frac{\partial}{\partial\beta} \alpha_{t-1}^\theta(v) + \alpha_{t-1}^\theta(v)\frac{\partial}{\partial\beta} B_{vs}) L_{t-1}(v). \tag{10.8}$$

$$\frac{\partial}{\partial\beta} \alpha_0^\theta(s) = \frac{\partial}{\partial\alpha} \alpha_0^\theta(s) = 0. \tag{10.9}$$

(10.10)

Because of the special structure of the transition matrix of a immigration model, we have the following.

$$B_{sv} = \mathbb{I}_{v=s+1} B_{s,s+1} + \mathbb{I}_{v=s} B_{ss} + \mathbb{I}_{v=s-1} B_{s,s-1}$$

$$\frac{\partial}{\partial\alpha} B_{vs} = (\mathbb{I}_{v=s+1} - \mathbb{I}_{v=s})(\frac{1}{\Omega} - \frac{\alpha}{\Omega^2}\frac{\partial\Omega}{\partial\alpha}) + (\mathbb{I}_{v=s-1} - \mathbb{I}_{v=s})(-\frac{\beta}{\Omega^2}\frac{\partial\Omega}{\partial\alpha} s) \tag{10.11}$$

$$\frac{\partial}{\partial\beta} B_{vs} = (\mathbb{I}_{v=s+1} - \mathbb{I}_{v=s})(-\frac{\alpha}{\Omega^2}\frac{\partial\Omega}{\partial\beta}) + (\mathbb{I}_{v=s-1} - \mathbb{I}_{v=s})(\frac{1}{\Omega} - \frac{\beta}{\Omega^2}\frac{\partial\Omega}{\partial\beta}) s \tag{10.12}$$

$$\Omega = max\{\beta\ ,\alpha\} + (d-2)\beta \tag{10.13}$$

$$\frac{\partial\Omega}{\partial\alpha} = k\mathbb{I}_{\alpha\geqslant\beta} \tag{10.14}$$

$$\frac{\partial\Omega}{\partial\beta} = k(d-2)\beta + \mathbb{I}_{\beta>\alpha} k \tag{10.15}$$

So we can get the derivative $\nabla_\theta P(y_{[T_0,T_{N+1})}|W,\theta)$ from FFBS algorithm.
Now briefly introduce Hamiltonian Monte Carlo.
Hamilton's Equations

$$
\begin{aligned}
\frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\
\frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i}
\end{aligned}
$$

, where H represents the total "energy", $H(q,p) = U(q) + K(p)$, and q is the "position" and p is the momentum.
$K(p) = p^\top \Sigma p$, where Σ is a positive definite symmetric matrix. Usually people use $K(p) = \sum_{i=1}^{d} \frac{p_i^2}{m_i}$.

Leapfrog Method:

$$
\begin{aligned}
p_i(t+\epsilon/2) &= p_i(t) - \frac{\epsilon}{2}\frac{\partial U}{q_i}(q(t)) \\
q_i(t+\epsilon) &= q_i(t) + \epsilon \frac{p_i(t+\epsilon/2)}{m_i} \\
p_i(t+\epsilon) &= p_i(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial U}{q_i}(q(t+\epsilon))
\end{aligned}
$$

For our case, $q = \theta = (\alpha, \beta)^\top)$, $p = (p_1, p_2)^\top$.

$$
\begin{aligned}
U(\theta) &= -\log(P(\theta|W,y)) \\
K(p) &= \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} \\
H(\theta,p) = P(\theta,p) = \frac{1}{Z}\exp(-U(\theta))\exp(-K(p)) &= \frac{1}{Z}P(\theta|W,y)e^{-\frac{p_1^2}{m_1}}e^{-\frac{p_2^2}{m_2}}.
\end{aligned}
$$

So, our

11. Beam Sampling for continuous time Infinite Hidden Markov Models. Model Assumptions:
We are considering a continuous time Infinite Hidden Markov Model (iHMM), with transition matrix $A = (A_{ij})_{i,j\geqslant 1}$.

$$
\begin{aligned}
A_i \doteq A_{ii} &= -\sum_{j\neq i} A_{ij} \\
A_{ij} &> 0 \; , i \neq j
\end{aligned}
$$

Now, we are using the way in reference 1 to construct a MJPs with virtual jumps.

PROPOSITION 11.1. *The path (W,V) returned by the thinning procedure described in algorithm 3 is equivalent to a sample (S,T) from the $MJP(\pi_0, A)$.*

Proof. $S = (s_0, s_1, ..., s_N)$,$T = (t_0, t_1, ..., t_N, t_{N+1})$. And let's call the virtual jumps as U. Denote the virtual jump times between (t_i, t_{i+1}) as n_i. Then the density

Algorithm 2 HMC In Gibbs sampling for MJPs

Input: observations $y_{[t_0,t_{k+1})}$
Initialize, $i = 0$
(a) Set $\theta(0)$ 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 $p_1^* \sim N(0, m_1)$, $p_2^* \sim N(0, m_2)$.
    (b) Start with the current state $(\theta(i-1), p_1^*, p_2^*)$, Hamilton dynamics is simulated for L steps using the leapfrog method with a stepsize ϵ to get (θ^*, p^{**}) then propose a new state as $(\theta^*, -p^{**})$,
    where

$$H(\theta, p) = P(\theta, p) = \frac{1}{Z}\exp(-U(\theta))\exp(-K(p)) = \frac{1}{Z}P(\theta|W,y)P(W,y)e^{-\frac{p_1^2}{m_1}}e^{-\frac{p_2^2}{m_2}}.$$

    (c) Set $\theta(i) = \theta^*$, with probability $\alpha = 1 \wedge \exp(-H(\theta^*, -p^{**}) + H(\theta^*, p^*))$;
    Otherwise set $\theta(i) = \theta(i-1)$.
    (d) Use FFBS algorithm to sample states given all the jump times(both true jumps and virtual jumps). (i.e. $V(i) \sim P(V|\theta(i), W, y)$.) Then delete all the virtual jumps to get $S(i), T(i)$.
    (e) Sample $U(i) \sim P(U|\theta(i), S(i), T(i), y)$.
  end for
until $i = N$

Algorithm 3 State-dependent thinning for MJPs

Input: Transition matrix $A_{ss'}$, and an initial distribution over states π_0.
Dominating Transition Rate Vector $B_s \geqslant A_s$.
Output: A piecewise constant trajectory $(V, W) = ((v_i, w_i))$ on the time interval $[t_{start}, t_{end}]$.
Initialize,
Draw $v_0 \sim \pi_0$ and set $w_0 = t_{start}$. Set $i = 0$.
while $w_i < t_{end}$ **do**
  (a) Sample $\tau_i \sim B_{v_i}$.
  (b) Set $v_{i+1} = v_i$ with probability $1 - \frac{A_{v_i}}{B_{v_i}}$ and set $w_{i+1} = w_i + \tau_i$.
  (c) **Else:** Set $w_{i+1} = w_i + \tau_i$ and sample v_{i+1} with $P(v_{i+1} = s|v_i) = A_{v_i,s}/A_{v_i}$.
  (d) Incresement i.
end while

function of (W, V) will be as follows.

$$\begin{aligned}P(W,V) &= \pi_0(s_0)\prod_{i=0}^{N-1}\exp(-B_{s_i}(t_{i+1}-t_i))B_{s_i}^{n_i}(1-\frac{A_{s_i}}{B_{s_i}})^{n_i}B_{s_i}\frac{A_{s_i}}{B_{s_i}}\cdot\exp(-B_{s_N}(t_{N+1}-t_N))B_{s_N}^{n_N}(1-\frac{A_{s_N}}{B_{s_N}})^{n_N} \\ &= \pi_0(s_0)\exp(-\int_{t_0}^{t_{N+1}}B_{S_{(t)}}dt)\prod_{i=0}^{N}(B_{s_i}-A_{s_i})^{n_i}\prod_{i=0}^{N-1}A_{s_is_{i+1}}\end{aligned}$$

So after integrating with respect to virtual jump times and the numbers of virtual jumps, we can get the following.

$$P(S,T) = \sum_{n_1,n_2,...,n_N \geqslant 0} \int_{t_0 \leqslant \tau_1^1 \leqslant ... \leqslant \tau_{n_1}^1 \leqslant t_1} ... \int_{t_N \leqslant \tau_1^N \leqslant ... \leqslant \tau_{n_N}^N \leqslant t_{N+1}} P(W,V) d\tau_1^1 ... d\tau_{n_1}^1 ... d\tau_1^N ... d\tau_{n_N}^N$$
$$= \pi_0(s_0) \exp(-\int_{t_0}^{t_{N+1}} A_{S_{(t)}} dt) \prod_{i=0}^{N-1} A_{s_i s_{i+1}}$$

So the proposition is proved.

□

The main idea of beam sampler for infinite-state continuous time Hidden Markov Model is to introduce auxiliary variables μ such that conditioned on μ, the number of trajectories with positive probability is finite. Then dynamic programming can be used to compute the conditional probabilities efficiently.

Assume $W = (w_0, w_1, ..., w_{N'}, w_{N'+1})$, $V = (v_0, v_1, ..., v_{N'})$, $\mu = (\mu_1, \mu_2, ..., \mu_{N'})$.

$$P(\mu|W,V) = \prod_{i=1}^{N'} \frac{\mathbb{I}(0 \leqslant \mu_i \leqslant \frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}})}{\frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}}}$$

It indicates that conditioned on the trajectory(V,W), μ_i is depending on A, v_i, and v_{i-1} and $\mu_i \sim Uniform(0, \frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}})$.

Proposition 11.2. *Conditioned on a trajectory (S,T) of the MJP, the virtual jump times U are distributed as a Poisson process with density $B_{s(t)} - A_{s(t)}$.*

Proof. $S = (s_0, s_1, ..., s_N)$,$T = (t_0, t_1, ..., t_N, t_{N+1})$. And let's call the virtual jumps as U. Denote the virtual jump times between (t_i, t_{i+1}) as n_i. Then the density function of (W,V) will be as follows.

$$P(W,V) = P(U,S,T)$$
$$= \pi_0(s_0) \exp(-\int_{t_0}^{t_{N+1}} B_{S_{(t)}} dt) \prod_{i=0}^{N} (B_{s_i} - A_{s_i})^{n_i} \prod_{i=0}^{N-1} A_{s_i s_{i+1}}$$

$$P(S,T,n_0,...,n_N) = \pi_0(s_0) \exp(-\int_{t_0}^{t_{N+1}} B_{S_{(t)}} dt) \prod_{i=0}^{N} \frac{((B_{s_i} - A_{s_i})(t_{i+1} - t_i))^{n_i}}{n_i!} \prod_{i=0}^{N-1} A_{s_i s_{i+1}}$$

So the conditional probability $P(n_0, n_1, ..., n_N|S,T)$ will be as follows.

$$P(n_0,...,n_N|S,T) = \exp(-\int_{t_0}^{t_{N+1}} (B_{S_{(t)}} - A_{S_{(t)}}) dt) \prod_{i=0}^{N} \frac{((B_{s_i} - A_{s_i})(t_{i+1} - t_i))^{n_i}}{n_i!}$$

So it indicates that conditioned on the trajectory (S, T), the virtual jump U is distributed as a non-homogeneous Poisson process with density $B_{s(t)} - A_{s(t)}$. $\square$

Sampling v: Using the same trick used in Beam Sampling for the Infinite HMM, we can sample $P(v_t|y, \mu, W)$. So can we sample $P(v_t|v_{t+1}, y, W, u)$.

First of all, consider $P(v_i|y_{w_0,w_{i+1}}, w_{0:i}, \mu_{0:i})$.

$$
\begin{aligned}
P(v_i, w_{0:i}, \mu_{0:i}, y_{[w_0,w_{i+1})}) &= \sum_{v_{i-1}} P(v_i, y_{[w_i,w_{i+1})}, w_i, \mu_i, v_{i-1}, w_{0:i-1}, \mu_{0:i-1}, y_{[w_0,w_i)}) \\
&= \sum_{v_{i-1}} P(v_i, y_{[w_i,w_{i+1})}, w_i, \mu_i | v_{i-1}, w_{0:i-1}, \mu_{0:i-1}, y_{[w_0,w_i)}) P(v_{i-1}, w_{0:i-1}, \mu_{0:i-1}, y_{[w_0,w_i)}) \\
&= \sum_{v_{i-1}} P(y_{[w_i,w_{i+1})}|v_i, w_i, w_{i+1}) P(\mu_i|v_i, v_{i-1}) P(v_i, w_i|v_{i-1}, w_{i-1}) P(v_{i-1}, w_{0:i-1}, \mu_{0:i-1}, y_{[w_0,w_i)}) \\
&= P(y_{[w_i,w_{i+1})}|v_i, w_i, w_{i+1}) \sum_{v_{i-1}} \frac{\mathbb{I}(0 \leqslant \mu_i \leqslant \frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}})}{\frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}}} \exp(-B_{v_{i-1}}(w_i - w_{i-1})) \\
&\quad (B_{v_{i-1}} - A_{v_{i-1}})^{\mathbb{I}(v_i = v_{i-1})} A_{v_{i-1}v_i}^{\mathbb{I}(v_i v_{i-1})} P(v_{i-1}, w_{0:i-1}, \mu_{0:i-1}, y_{[w_0,w_i)}) \\
&= P(y_{[w_i,w_{i+1})}|v_i, w_i, w_{i+1}) \sum_{\Im_{i-1}} \frac{\mathbb{I}(0 \leqslant \mu_i \leqslant \frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}})}{\frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}}} \exp(-B_{v_{i-1}}(w_i - w_{i-1})) \\
&\quad (B_{v_{i-1}} - A_{v_{i-1}})^{\mathbb{I}(v_i = v_{i-1})} A_{v_{i-1}v_i}^{\mathbb{I}(v_i v_{i-1})} P(v_{i-1}, w_{0:i-1}, \mu_{0:i-1}, y_{[w_0,w_i)})
\end{aligned}
$$

Although the summation over v_{i-1} is an infinite sum, the auxiliary variable μ_i truncates this summation to the finitely many v_{i-1}'s and v_i's that satisfy both constrains $\mu_i \leqslant \frac{A_{v_{i-1}v_i}}{A_{v_{i-1}}}$ and $P(v_{i-1}|y_{[w_0,w_i)}, \mu_{0:i-1}) > 0$. This means that $|\Im_{i-1}| < +\infty$.

Secondly, consider $P(v_i|v_{i+1}, y_{w_0,w_{N'+1}}, w_{0:N'+1}, \mu_{0:N'})$.

$$
\begin{aligned}
&P(v_i|v_{i+1}, y_{w_0,w_{N'+1}}, w_{0:N'+1}, \mu_{0:N'}) \propto P(v_i, v_{i+1}, y_{w_0,w_{N'+1}}, w_{0:N'+1}, \mu_{0:N'}) \\
&= P(y_{[w_{i+1},w_{N'+1})}, \mu_{i+2:N}, w_{i+2:N}|v_i, v_{i+1}, y_{[w_0,w_{i+1}),w_{0:i+1},\mu_{0:i+1})}) P(v_i, v_{i+1}, y_{[w_0,w_{i+1})}, w_{0:i+1}, \mu_{0:i+1}) \\
&= P(y_{[w_{i+1},w_{N'+1})}, \mu_{i+2:N}, w_{i+2:N}|v_{i+1}, w_{i+1}) P(v_i, v_{i+1}, y_{[w_0,w_{i+1})}, w_{0:i+1}, \mu_{0:i+1}) \\
&= Const \cdot P(v_i, v_{i+1}, y_{[w_0,w_{i+1})}, w_{0:i+1}, \mu_{0:i+1}) \\
&= Const \cdot P(v_{i+1}, u_{i+1}, w_{i+1}|v_i, y_{[w_0,w_{i+1})}, w_{0:i}, \mu_{0:i}) \cdot P(v_i, w_{0:i}, \mu_{0:i}, y_{[w_0,w_{i+1})}) \\
&= Const \cdot P(v_{i+1}, u_{i+1}, w_{i+1}|v_i, w_{0:i}, \mu_{0:i}) \cdot P(v_i, w_{0:i}, \mu_{0:i}, y_{[w_0,w_{i+1})})
\end{aligned}
$$

Finally, to sample the complete trajectory, we can sample $P(v_{N'}|y_{w_0,w_{N'+1}}, \mu_{0:N'})$ first, and then do a backward sampling using the above formula.

THEOREM 11.3. *Algorithm 4 has $P(S, T, W, \mu|y)$ as a stationary distribution.*

Proof. Firstly, prove (c) step has $P(S, T|W, \mu, y)$ as a stationary distribution. It comes from the following detail balance condition.

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

Algorithm 4 Beam Sampler for continuous time Infinite Hidden Markov Models

Input: observations $y_{[t_0,t_{k+1})}$, A, B, π_0
Initialize, $i=0$
(a) Set current trajectory $[S,T](0)$ arbitrarily.
repeat
  for $i=0$ **to** N **do**
    (a) Sample virtual jumps $U(i+1) \sim \mathit{Poisson\ Process}(B_{s(t)} - A_{s(t)})$, given $S(i), T(i)$.
    (b) Sample $\mu(i+1)_j \sim \mathit{Uniform}(0, \frac{A_{v_{j-1}v_j}}{A_{v_{j-1}}})$, $j=1,2,...,N'$.
    (c) Sample $V(i+1) \sim P(V|W(i+1), \mu(i+1), y)$
    (d) Delete all the virtual jumps to get $S(i+1)$, $T(i+1)$
  end for
until $i=N$

Secondly, prove (a) and (b) step have $P(W,\mu|S,T,y)$ as a stationary distribution.

$$\begin{aligned}
P(W,\mu|S,T,y) &= P(U,\mu|S,T,y) = \frac{P(U,\mu,S,T,y)}{P(S,T,y)} \\
&= \frac{P(y|S,T)P(U,\mu,S,T)}{P(y|S,T)P(S,T)} = P(\mu|S,T,U)P(U|S,T) \\
&= P(\mu|V,W)P(U|S,T)
\end{aligned}$$

We know the transition probability $P((S,T,W,\mu) \to (S,T,W^*,\mu^*))$ is as follows.

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

So step(a) and (b) have $P(W,\mu|S,T,y)$ as a stationary distribution.
Above all, this theorem is proved.
□

12. Delayed Acceptance MH algorithm for MJPs.

Proof. First prove that Step (a) - (d) have $P(\theta|W,y)P(S_a,T_a|W,\theta)$ as stationary distribution.
From step(a) to step (d), W, y, S, T stay unchanged.
Let the first stage acceptance rate

$$\begin{aligned}
\alpha_1(. \to *) &= \alpha_1(S_a,T_a,\theta) \to (S_a^*,T_a^*,\theta^*) \\
&\doteq 1 \wedge \frac{P(y|S_a^*,T_a^*,\theta^*)q(\theta|\theta^*)}{P(y|S_a,T_a,\theta)q(\theta^*|\theta)}.
\end{aligned}$$

Let the second stage acceptance rate

$$\begin{aligned}
\alpha_2(. \to *) &= \alpha_2(S_a,T_a,\theta) \to (S_a^*,T_a^*,\theta^*) \\
&\doteq 1 \wedge \frac{P(\theta^*|W,y)P(y|S_a,T_a,\theta)}{P(\theta|W,y)P(y|S_a^*,T_a^*,\theta^*)}.
\end{aligned}$$

Algorithm 5 Delayed Acceptance MH algorithm for MJPs

Input: observations $y_{[t_0,t_{k+1})}$
Initialize, $i = 0$
(a) Set $\theta(0)$ arbitrarily and set current trajectory $[S,T](0)$ arbitrarily.
(b) Sample virtual jumps $U(0)$ based on $[S,T](0)$.
repeat
  for $i = 1$ **to** N **do**
    (a) Propose $\theta^* \sim q(\theta^*|\theta)$.
    (b) Sample the adjoint trajectory $S_a^*, T_a^* \sim P(S,T|W(i-1),\theta^*)$ for proposed θ^*. Sample the adjoint trajectory $S_a, T_a \sim P(S,T|W(i-1),\theta(i-1))$
    (c) With probability

$$1 \wedge \frac{P(y|S_a^*,T_a^*,W(i-1),\theta^*)}{P(y|S_a,T_a,W(i-1),\theta(i-1))}\frac{q(\theta(i-1)|\theta^*)}{q(\theta^*|\theta(i-1))}$$

    Run the following Forward Filter algorithm. Otherwise, set $\theta(i) = \theta(i-1)$, then increase i and go to (e).
    (d) With probability

$$1 \wedge \frac{P(\theta^*|W(i-1),y)}{P(\theta|W(i-1),y)}\frac{P(y|S_a,T_a,W(i-1),\theta(i-1))}{P(y|S_a^*,T_a^*,W(i-1),\theta^*)}$$

    Set $\theta(i) = \theta^*$. Otherwise, set $\theta(i) = \theta(i-1)$.
    (e) Use FFBS algorithm to sample states given all the jump times(both true jumps and virtual jumps). (i.e. $V(i) \sim P(V|\theta(i),W(i-1),y)$.) Then delete all the virtual jumps to get $S(i), T(i)$.
    (f) Sample $U(i) \sim P(U|\theta(i),S(i),T(i),y)$.
  end for
until $i = N$

We know the transition probability is as follows.

$$P((S_a,T_a,\theta) \to (S_a^*,T_a^*,\theta^*)) = q(\theta^*|\theta)P(S_a^*,T_a^*|W,\theta^*)\alpha_1(. \to *)\alpha_2(. \to *).$$

First of all, we have,

$$\begin{aligned}P(y|S_a,T_a,W,\theta)q(\theta^*|\theta)\alpha_1(. \to *) &= P(y|S_a,T_a,W,\theta)q(\theta^*|\theta)\alpha(. \to *) \wedge P(y|S_a^*,T_a^*,W,\theta^*)q(\theta|\theta^*)\\ &= P(y|S_a^*,T_a^*,W,\theta^*)q(\theta|\theta^*)\alpha_1(* \to .).\end{aligned}$$

Secondly, we have,

$$\begin{aligned}&\frac{P(\theta|W,y)P(S_a,T_a|W,\theta)P(S_a^*,T_a^*|W,\theta^*)}{P(y|S_a,T_a,W,\theta)}\alpha_2(. \to *)\\ &= P(S_a,T_a|W,\theta)P(S_a^*,T_a^*|W,\theta^*)\left(\frac{P(\theta|W,y)}{P(y|S_a,T_a,W,\theta)} \wedge \frac{P(\theta^*|W,y)}{P(y|S_a^*,T_a^*,W,\theta^*)}\right)\\ &= \frac{P(\theta^*|W,y)P(S_a^*,T_a^*|W,\theta^*)P(S_a,T_a|W,\theta)}{P(y|S_a^*,T_a^*,W,\theta^*)}\alpha_2(* \to .).\end{aligned}$$

Algorithm 6 New MH algorithm for MJPs

Input: observations $y_{[t_0,t_{k+1})}$
Initialize, $i = 0$
(a) Set $\theta(0)$ arbitrarily and set current trajectory $[S,T](0)$ arbitrarily.
(b) Sample virtual jumps $U(0)$ based on $[S,T](0)$.
repeat
  for $i = 1$ **to** N **do**
    (a) Propose $\theta^* \sim q(\theta^*|\theta)$.
    (b) Sample the adjoint trajectory $S_a^*, T_a^* \sim P(S,T|W(i-1),\theta^*)$ for proposed θ^*. Sample the adjoint trajectory $S_a, T_a \sim P(S,T|W(i-1),\theta(i-1))$
    (c) With probability

$$1 \wedge \frac{P(y|S_a^*,T_a^*,W(i-1),\theta^*)}{P(y|S_a^*,T_a^*,W(i-1),\theta(i-1)}\frac{P(\theta^*)P(W(i-1)|\theta^*)}{P(\theta)P(W(i-1)|\theta(i-1))}$$

    Set $\theta(i) = \theta^*$. Otherwise, set $\theta(i) = \theta(i-1)$.
    (d) Use FFBS algorithm to sample states given all the jump times(both true jumps and virtual jumps). (i.e. $V(i) \sim P(V|\theta(i),W(i-1),y)$.) Then delete all the virtual jumps to get $S(i),T(i)$.
    (e) Sample $U(i) \sim P(U|\theta(i),S(i),T(i),y)$.
  end for
until $i = N$

So above all, we have,

$$\begin{aligned}
&P(\theta|W,y)P(S_a,T_a|\theta,W)P((S_a,T_a,\theta)\to(S_a^*,T_a^*,\theta^*))\\
&= P(\theta|W,y)P(S_a,T_a|\theta,W)q(\theta^*|\theta)\alpha_1(.\to *)\alpha_2(.\to *)P(S_a^*,T_a^*|W,\theta^*)\\
&= P(y|S_a,T_a,W,\theta)\alpha_1(.\to *)q(\theta^*|\theta)\cdot\frac{P(\theta|W,y)P(S_a,T_a|W,\theta)P(S_a^*,T_a^*|W,\theta^*)}{P(y|S_a,T_a,W,\theta)}\alpha_2(.\to *)\\
&= P(y|S_a^*,T_a^*,W,\theta^*)\alpha_1(*\to .)q(\theta|\theta^*)\cdot\frac{P(\theta^*|W,y)P(S_a^*,T_a^*|W,\theta^*)P(S_a,T_a|W,\theta)}{P(y|S_a^*,T_a^*,W,\theta^*)}\alpha_2(*\to .)\\
&= P(\theta^*|W,y)P(S_a^*,T_a^*|\theta^*,W)q(\theta|\theta^*)\alpha_1(*\to .)\alpha_2(*\to .)P(S_a,T_a|W,\theta)\\
&= P(\theta^*|W,y)P(S_a^*,T_a^*|\theta^*,W)P((S_a^*,T_a^*,\theta^*)\to(S_a,T_a,\theta))
\end{aligned}$$

So Step(a) - Step(d) $P(\theta|W,y)P(S_a,T_a|\theta)$ as stationary distribution. So if we only keep θ, then θs are distributed as $P(\theta|W,y) = \sum_{S_a,T_a} P(\theta|W,y)P(S_a,T_a,|W,\theta)$.
Then the following proof will be exactly the same as the algorithm 1 (MH In Gibbs sampling for MJPs). ∎

13. New MH algorithm for MJPs.

Proof. First prove that Step (a) - (c) have $P(y|W,\theta,S_a,T_a)P(S_a,T_a|W,\theta)P(W,\theta)$ as stationary distribution.
From step(a) to step (c), W, y, S, T stay unchanged.
Step (a) - (c) is exactly a pseudo marginal MH scheme.
∎

14. Variance Analysis on MH sampler APR 21. In this section, we consider two conditional variances, $Var(\beta|S,T)$ and $Var(\beta|W,y)$. If the first one is smaller, it means that the information provided from S,T is more than the information

provided from W. Then we should prefer the MH sampler instead of the Gibbs sampler. Vice versa.
Since given S, T, β is distributed as $Gamma(\omega + D, \sum_{i=0}^{d-1} \tau_i i + \theta)$, the conditional variance

$$Var(\beta|S,T) = \frac{\omega + D}{(\sum_{i=0}^{d-1} \tau_i i + \theta)^2}.$$

For immigration model, if $\alpha < \beta$, then the rate of W, $\Omega = k(d-1)\beta$.
Then $\beta|W, \alpha < \beta \sim Gamma(\omega + |W|, k(d-1)(T_{N+1} - T_0) + \theta)$. So the conditional variance $Var(\beta|W, \alpha < \beta)$ will be as follows.

$$Var(\beta|W, \alpha < \beta) = \frac{\omega + |W|}{(k(d-1)(T_{N+1} - T_0) + \theta)^2}.$$

It implies that the second conditional variance is smaller, which means we should choose the Gibbs sampler instead of the MH sampler.

15. Generic Metropolis Hasting using FFBS within the Gibbs Sampling On MJPs MAY1.
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 β, S, T, α is distributed as a $Gamma$ distribution.
$\alpha|\beta, S, T, y = \alpha|\beta, S, T \sim Gamma(\mu + N, \lambda + \sum_0^N F_{S_i}(\beta)(t_{i+1} - t_i))$.

Algorithm 7 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 $i = N$

Algorithm 8 Generic MH In 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) Propose $\beta^* \sim q(.|\beta(i-1))$.
    Set $\beta(i) = \beta^*$, with probability $\alpha = 1 \wedge \frac{P(\beta^*|W(i-1),\alpha(i-1),y)}{P(\beta(i-1)|W(i-1),\alpha(i-1),y)} \frac{q(\beta(i-1)|\beta^*)}{q(\beta^*|\beta(i-1))}$;
    Otherwise set $\beta(i) = \beta(i-1)$.
    The acceptance probability is as follows.
    $\alpha = 1 \wedge \frac{P(y|W(i-1),\alpha(i-1),\beta^*)p_2(\beta^*)}{P(y|W(i-1),\alpha(i-1),\beta(i-1))p_2(\beta)} \frac{q(\beta(i-1)|\beta^*)}{q(\beta^*|\beta(i-1))}$
    (b) Use FFBS algorithm to sample states given all the jump times(both true jumps and virtual jumps). (i.e. $V(i) \sim P(V|\beta(i), \alpha(i-1), W(i-1), y)$.) Then delete all the virtual jumps to get $S(i), T(i)$.
    (c) Sample $U(i) \sim P(U|\beta(i), \alpha(i-1), S(i), T(i), y)$.
    (d) Sample $\alpha(i) \sim P(.|\beta(i), S(i), T(i), y)$. It is a Gamma distribution actually.
  end for
until $i = N$

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.

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

Now, consider a MH sampler for such models.

Algorithm 9 Revised Generic MH In 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)
(i) Propose $\beta^* \sim q(.|\beta(i-1))$.
(ii) Set $\Omega = max\{F_\Omega(\alpha(i-1), \beta(i-1)), F_\Omega(\alpha(i-1), \beta^*)\}$.
(b) Sample $U(i) \sim P_\Omega(U|\beta(i), \alpha(i-1), S(i), T(i), y)$.
(c) Set $\beta(i) = \beta^*$, with probability

$$\alpha = 1 \wedge \frac{P_\Omega(\beta^*|W(i-1), \alpha(i-1), y)}{P_\Omega(\beta(i-1)|W(i-1), \alpha(i-1), y)} \frac{q(\beta(i-1)|\beta^*)}{q(\beta^*|\beta(i-1))};$$

Otherwise set $\beta(i) = \beta(i-1)$.
The acceptance probability is as follows.
$\alpha = 1 \wedge \frac{P_\Omega(y|W(i-1),\alpha(i-1),\beta^*)p_2(\beta^*)}{P_\Omega(y|W(i-1),\alpha(i-1),\beta(i-1))p_2(\beta)} \frac{q(\beta(i-1)|\beta^*)}{q(\beta^*|\beta(i-1))}$
(d) Use FFBS algorithm to sample states given all the jump times(both true jumps and virtual jumps). (i.e. $V(i) \sim P(V|\beta(i), \alpha(i-1), W(i-1), y)$.) Then delete all the virtual jumps to get $S(i), T(i)$.
(e) Sample $\alpha(i) \sim P(.|\beta(i), S(i), T(i), y)$. It is a Gamma distribution actually.
end for
until $i = N$

16. Generic Metropolis Hasting using FFBS within the Gibbs Sampling On MJPs AUG 5.

17. Figures and tables.

18. Bibliography and BibTEX.

19. Conclusion. Appendix. The use of appendices.

Appendix A. Title of appendix.

REFERENCES

[1] VINAYAK RAO, YEE WHYE TEH, *MCMC for continuous-time discrete-state systems*, NIPS, 2012.