Supplement to 'Inhomogeneous hidden semi-Markov models for incompletely observed point processes'

Amina Shahzadi*[ab], Ting Wang †[a], Mark Bebbington[c], and Matthew Parry[a]

[a]Department of Mathematics and Statistics, The University of Otago, Dunedin, New Zealand.
[b]Department of Statistics, Government College University, Lahore, Pakistan.
[c]Volcanic Risk Solutions, Massey University, Palmerston North, New Zealand.

June 14, 2022

S.1 The forward-backward algorithm

The forward-backward algorithm for hidden Markov models (HMMs) was originally introduced by Baum et al (1970) whose extended formulation for various hidden semi-Markov models (HSMMs) can be found in Ferguson (1980), Levinson (1986), Murphy (2002), Guédon (2003), Bulla (2006) and Yu (2010). Since, our proposed inhomogeneous hidden semi-Markov model (IHSMM) is based on the right-censored assumption of time-spent in the last visited state proposed by Guédon (2003), we extend the forward-backward algorithm with the above assumption for the proposed IHSMM.

In HSMMs, the conditional independence assumption between past and future of the process is satisfied at the times of a state change of the hidden semi-Markov chain. With this key assumption, the forward-backward algorithm in HSMMs is based on the following decomposition:

$$\begin{aligned}\mathcal{P}_t(j) &= \Pr\left(X_1, \ldots, X_T, S_t = j, S_{t+1} \neq j \,|\, \Theta\right) \\ &= \Pr\left(X_1, \ldots, X_t, S_t = j, S_{t+1} \neq j \,|\, \Theta\right) \Pr\left(X_{t+1}, \ldots, X_T \,|\, S_t = j, S_{t+1} \neq j, \Theta\right) \\ &= \alpha_t(j)\beta_t(j), \end{aligned} \tag{1}$$

where

$$\alpha_t(j) = \Pr\left(X_1, \ldots, X_t, S_t = j, S_{t+1} \neq j \,|\, \Theta\right) \tag{2}$$

is the forward probability which is defined as the joint probability of the occurrence of the first t observations with the tth observation being the last of a sequence from state j given the model parameter Θ, and

$$\beta_t(j) = \Pr\left(X_{t+1}, \ldots, X_T \,|\, S_t = j, S_{t+1} \neq j, \Theta\right) \tag{3}$$

is the backward probability which is defined as the conditional probability of the partial observation sequence $X_{t+1}, \ldots, X_T$ given that state j ended at time t and the model parameter Θ.

*amina@maths.otago.ac.nz
†ting.wang@otago.ac.nz

S.1.1 The forward algorithm

Borrowing the definition of the forward algorithm for an HSMM in (2), the recursive forward variable for an IHSMM with m states is given by

$$\begin{aligned}\alpha_t(j) &= \Pr\left(X_1, \ldots, X_t; S_t = j, S_{t+1} \neq j \mid \Theta\right) \\ &= \pi_j p_j(t, 1) \prod_{l=1}^{t} f_j(x_l) + \sum_{\substack{i=1 \\ i\neq j}}^{m} \sum_{d=1}^{t-1} \alpha_{t-d}(i) a_{ij}\, p_j(d, t-d+1) \prod_{l=t-d+1}^{t} f_j(x_l), \end{aligned} \tag{4}$$

for $t = 1, \ldots, T-1$ and $j = 1, \ldots, m$, with the second term on the right hand side being zero when $t = 1$. When $t = T$, the sojourn time in the last visited state is right censored. Since, only the minimum time spent in the last visited is known, the probability distribution of durations of state j, $p_j(d, T)$, is replaced by the corresponding survival function $\bar{P}_j(d, T) = \sum_{k \geqslant d} p_j(k, T)$ in the last term of the general forward recursion formula in (4). Therefore, we have

$$\begin{aligned}\alpha_T(j) &= \Pr\left(X_1, \ldots, X_T; S_T = j \mid \Theta\right) \\ &= \pi_j \bar{P}_j(T, 1) \prod_{l=1}^{T} f_j(x_l) + \sum_{\substack{i=1 \\ i\neq j}}^{m} \sum_{d=1}^{T-1} \alpha_{T-d}(i) a_{ij}\, \bar{P}_j(d, T-d+1) \prod_{l=T-d+1}^{T} f_j(x_l), \end{aligned} \tag{5}$$

for $j = 1, \ldots, m$. To reduce the computational work, the product of state-dependent probabilities for state j can be computed by using the recursive formula proposed by Mitchell and Jamieson (1993) as below. Let

$$u_t(j, d) = \prod_{l=t-d+1}^{t} f_j(x_l), \tag{6}$$

then, we can write

$$u_t(j, d) = u_{t-1}(j, d-1) f_j(x_t), \qquad \text{for } t > 1 \text{ and } d > 1, \tag{7}$$

with $u_t(j, 1) = f_j(x_t)$ for $t = 1, \ldots, T$. Now we introduce a new variable $\alpha^*_{t+1}(j)$ which is useful in evaluating the forward recursion, and is defined by

$$\begin{aligned}\alpha^*_{t+1}(j) &= \Pr(X_1, \ldots, X_t,\ S_{t+1} = j,\ S_t \neq j \mid \Theta) \\ &= \sum_{\substack{i=1 \\ i\neq j}}^{m} \Pr(X_1, \ldots, X_t,\ S_{t+1} = j,\ S_t = i \mid \Theta) \\ &= \sum_{\substack{i=1 \\ i\neq j}}^{m} \alpha_t(i) a_{ij}, \end{aligned} \tag{8}$$

for $t = 1, \ldots, T-1$ with $\alpha^*_1(j) = \pi_j$. Here, $\alpha^*_{t+1}(j)$ expresses the joint probability of the partial observation sequence $X_{1:t}$ and state j commences at time $t+1$ given the model parameters Θ. Using (8) and (6) in (4) and (5), we have

$$\alpha_t(j) = \sum_{d=1}^{t} \alpha^*_{t-d+1}(j)\, p_j(d, t-d+1) u_t(j, d), \tag{9}$$

and

$$\alpha_T(j) = \sum_{d=1}^{T} \alpha^*_{T-d+1}(j)\, \bar{P}_j(d,\, T-d+1) u_T(j,\, d). \tag{10}$$

The likelihood function in (13) in the main manuscript can be evaluated as $\mathcal{L}(\Theta) = \sum_{j=1}^{m} \alpha_T(j)$.

S.1.2 The backward algorithm

From (3), the backward variable $\beta_t(j)$ is

$$\beta_t(j) = \Pr(X_{(t+1)} \ldots,\, X_T \,|\, S_t = j,\, S_{t+1} \neq j;\, \Theta),$$

which holds for $t = 1, \ldots, T-1$. We assume that the last visited state does not finish at time $t = T$, therefore, for $t = T-1, \ldots, 1$, we have

$$\begin{aligned}
\beta_t(j) &= \Pr\left(X_{(t+1):T} \,|\, S_t = j,\, S_{t+1} \neq j;\, \Theta\right) \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \sum_{d=1}^{T-t-1} \Pr\left(X_{(t+1):T},\, S_{(t+1):(t+d)} = i,\, S_{t+d+1} \neq i \,|\, S_t = j;\, \Theta\right) \\
&\qquad + \sum_{\substack{i=1 \\ i\neq j}}^{m} \Pr\left(X_{(t+1):T},\, S_{(t+1):T} = i \,|\, S_t = j,\, \Theta\right) \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \sum_{d=1}^{T-t-1} a_{ji}\, p_i(d,\, t+1) \prod_{l=t+1}^{t+d} f_i(x_l) \beta_{t+d}(i) + \sum_{\substack{i=1 \\ i\neq j}}^{m} a_{ji}\, \bar{P}_i(T-t,\, t+1) \prod_{l=t+1}^{T} f_i(x_l) \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \sum_{d=1}^{T-t-1} a_{ji}\, p_i(d,\, t+1) u_{t+d}(i,\, d) \beta_{t+d}(i) + \sum_{\substack{i=1 \\ i\neq j}}^{m} a_{ji}\, \bar{P}_i(T-t,\, t+1) u_T(i,\, T-t),
\end{aligned} \tag{11}$$

where $u_t(j,\, d)$ is defined in (6). The first term in (11) is zero when $t = T-1$ and the second term contributes to the right censoring of sojourn time in the last visited state. In the above backward recursion, $p_i(d,\, t+1)$ represents the probability distribution of durations for state i which depends on time entering in that state, i.e., $t+1$.

To make computational steps easier, we introduce another variable $\beta^*_{t+1}(j)$ defined by

$$\beta^*_{t+1}(j) = \Pr\left(X_{(t+1):T} \,|\, S_{t+1} = j,\, S_t \neq j\right), \tag{12}$$

which states the conditional probability of the partial observation sequence $X_{(t+1):T}$ given that state j starts at time $t+1$. By induction, we have

$$\begin{aligned}
\beta^*_{t+1}(j) &= \sum_{d=1}^{T-t-1} p_j(d,\, t+1) \prod_{l=t+1}^{t+d} f_j(x_l) \beta_{t+d}(j) + \bar{P}_j(T-t,\, t+1) \prod_{l=t+1}^{T} f_j(x_l) \\
&= \sum_{d=1}^{T-t-1} \beta_{t+d}(j)\, p_j(d,\, t+1) u_{t+d}(j,\, d) + \bar{P}_j(T-t,\, t+1)\, u_T(j,\, T-t).
\end{aligned} \tag{13}$$

Therefore, the relationship between $\beta_t(j)$ and $\beta^*_{t+1}(j)$ for $t = T-1, \ldots, 1$ is

$$\beta_t(i) = \sum_{\substack{j=1 \\ i \neq j}}^{m} a_{ij} \beta^*_{t+1}(j). \tag{14}$$

S.2 The conditional distribution of S_t

The conditional probability of state j at time t given the observation sequence $X_{1:T}$ is also referred to as posterior or smoothed probability. It is usually denoted by $\gamma_t(j)$, and is defined as

$$\gamma_t(j) = \Pr\left(S_t = j \,|\, X_{1:T};\ \Theta\right). \tag{15}$$

It can be easily computed using the forward-backward algortihm. Following Guédon (2003) the above quantity can be rewritten as follows.

$$\begin{aligned}
\gamma_t(j) &= \Pr\left(S_t = j \,|\, X_{1:T};\ \Theta\right) \\
&= \Pr\left(S_{t+1} = j \,|\, X_{1:T};\ \Theta\right) + \Pr\left(S_{t+1} \neq j,\, S_t = j \,|\, X_{1:T};\ \Theta\right) - \Pr\left(S_{t+1} = j,\, S_t \neq j \,|\, X_{1:T};\ \Theta\right) \\
&= \gamma_{t+1}(j) + \Pr\left(S_{t+1} \neq j,\, S_t = j \,|\, X_{1:T};\ \Theta\right) - \Pr\left(S_{t+1} = j,\, S_t \neq j \,|\, X_{1:T};\ \Theta\right).
\end{aligned} \tag{16}$$

The second and third term can be evaluated as:

$$\begin{aligned}
\Pr\left(S_{t+1} \neq j,\, S_t = j \,|\, X_{1:T};\, \Theta\right) &= \frac{\Pr\left(S_{t+1} \neq j,\, S_t = j,\, X_{1:T} \,|\, \Theta\right)}{\Pr\left(X_{1:T} \,|\, \Theta\right)} \\
&= \frac{\alpha_t(j)\beta_t(j)}{\mathcal{L}(\Theta)},
\end{aligned} \tag{17}$$

$$\begin{aligned}
\Pr\left(S_{t+1} = j,\, S_t \neq j \,|\, X_{1:T};\, \Theta\right) &= \frac{\Pr\left(S_{t+1} = j,\, S_t \neq j,\, X_{1:T} \,|\, \Theta\right)}{\Pr\left(X_{1:T} \,|\, \Theta\right)} \\
&= \frac{\alpha^*_{t+1}(j)\beta^*_{t+1}(j)}{\mathcal{L}(\Theta)}.
\end{aligned} \tag{18}$$

Thus, (16) becomes

$$\gamma_t(j) = \gamma_{t+1}(j) + \frac{\alpha_t(j)\beta_t(j)}{\mathcal{L}(\Theta)} - \frac{\alpha^*_{t+1}(j)\beta^*_{t+1}(j)}{\mathcal{L}(\Theta)}. \tag{19}$$

It can be evaluated recursively along with the backward variable through the backward pass with the initial condition at $t = T$ given by

$$\gamma_T(j) = \frac{\Pr\left(S_T = j,\, X_{1:T} \,|\, \Theta\right)}{\mathcal{L}(\Theta)} = \frac{\alpha_T(j)}{\sum_{j=1}^{m} \alpha_T(j)}. \tag{20}$$

Note that $\sum_{j=1}^{m} \gamma_t(j) = 1$ for each t.

S.3 Numerical issues in computing forward and backward variables

The forward and backward variables need to be scaled to address the problem of underflow because of the multiplication of probabilities. The paradigm of the conventional scaling of the

forward variable in HMMs is not appropriate for HSMMs. Guédon (2003) proposed a forward-backward algorithm with an embedded scaling for the right-censored HSMMs which is free from the numerical underflow problem and its computational complexity is quadratic in the worst case. A similar kind of scaling approach has also been considered by Murphy (2002) and was later proposed by Li and Yu (2015) in terms of robust scaling, coincidently. As we are using the forward-backward definitions from Ferguson (1980) which are severely affected by the underflow problem, we, therefore, implement the scaling procedure as proposed by Guédon (2003) and Li and Yu (2015) to control the numerical problem. The implementation of such scaling procedure will eventually make the scaled forward-backward algorithms equivalent to Guedon's forward-backward algorithm with the same amount of computational complexity, that is, $\mathcal{O}\left(mT(m+T)\right)$.

For scaling purposes, we introduce two symbols '$\check{}$' and '$\hat{}$'. Let c_t be the scaling factor at time t, then the symbol '$\check{}$' on a recursive variable, say $\check{\alpha}_t$, means that α_t is multiplied by $t-1$ scaling factors, that is, $\check{\alpha}_t = \prod_{l=1}^{t-1} c_l \alpha_t$. The recursive variable with symbol '$\hat{}$' on indicates that α_t is multiplied by t scaling factors, that is, $\widehat{\alpha}_t = \prod_{l=1}^{t} c_l \alpha_t = c_t \check{\alpha}_t$.

We define a variable $c_1^*(j)$ similar to $\alpha_1(j)$ in (9) for $t=1$ with the assumption that state j does not necessarily finish at time $t=1$,

$$\begin{aligned} c_1^*(j) &= \Pr\left(X_1, S_1 = j \,|\, \Theta\right) \\ &= \sum_{\substack{i=1 \\ i \neq j}}^{m} \alpha_1^*(j) \bar{P}_j(1,\,1) u_1(j,\,1). \end{aligned} \tag{21}$$

Then, the scaling factor c_1 at time $t=1$ is given by

$$c_1 = 1 \Big/ \sum_{j=1}^{m} c_1^*(j)\,,$$

We assume that

$$\widehat{\alpha}_1^*(j) = \alpha_1^*(j) = \pi_j, \tag{22}$$

then from (9), the scaled forward variable denoted by $\widehat{\alpha}_1(j)$ at $t=1$ is given by

$$\begin{aligned} \widehat{\alpha}_1(j) &= c_1\, \alpha_1(j) = \widehat{\alpha}_1^*(j) p_j(1,\,1)\, c_1\, u_1(j,\,1) \\ &= \widehat{\alpha}_1^*(j) p_j(1,\,1)\, \widehat{u}_1(j,\,1), \end{aligned}$$

where $\widehat{u}_1(j,\,1) = c_1 f_j(x_1)$ is the scaled observation probability at $t=1$. When $c_1, \ldots, c_{t-1}$ are known, then the scaling factor at time $t = 2, \ldots, T$, can be calculated by the following variable

$$\begin{aligned} c_t^*(j) &= \Pr\left(X_{1:t},\, S_t = j \,|\, \Theta\right) \prod_{l=1}^{t-1} c_l \\ &= \sum_{d=1}^{t} \alpha_{t-d+1}^*(j) \bar{P}_j(d,\, t-d+1) u_t(j,\, d) \prod_{l=1}^{t-1} c_l \\ &= \sum_{d=1}^{t} \prod_{k=1}^{t-d} c_k\, \alpha_{t-d+1}^*(j) \bar{P}_j(d,\, t-d+1) \prod_{l=t-d+1}^{t-1} c_l\, u_t(j,\, d) \\ &= \sum_{d=1}^{t} \widehat{\alpha}_{t-d+1}^*(j) \bar{P}_j(d,\, t-d+1) \check{u}_t(j,\, d), \end{aligned} \tag{23}$$

where using (6) and (7), $\breve{u}_t(j,\, d)$ is defined by

$$\begin{aligned}\breve{u}_t(j,\, d) &= \prod_{l=t-d+1}^{t-1} c_l\, u_t(j,\, d) \\ &= \prod_{l=t-d+1}^{t-1} c_l\, u_{t-1}(j,\, d-1) f_j(x_t) \\ &= c_{t-1} \prod_{l=t-d+1}^{t-2} c_l\, u_{t-1}(j,\, d-1) f_j(x_t) \\ &= c_{t-1}\, \breve{u}_{t-1}(j,\, d-1) f_j(x_t) \quad \text{for } d > 1, \end{aligned} \tag{24}$$

and

$$\breve{u}_t(j,\, d) = f_j(x_t) \quad \text{for } d = 1. \tag{25}$$

Thus, the scaling factor at time $t = 2, \ldots, T$ is

$$c_t = 1 \Big/ \sum_{j=1}^{m} c_t^*(j)\,, \tag{26}$$

and the scaled forward variable $\widehat{\alpha}_t(j)$ is

$$\begin{aligned}\widehat{\alpha}_t(j) &= \prod_{l=1}^{t} c_l \alpha_t(j) \\ &= c_t \prod_{l=1}^{t-1} c_l \alpha_t(j) \\ &= c_t \breve{\alpha}_t(j), \end{aligned} \tag{27}$$

where using equations (9), (10), (23) and (26), for $t = 1, \ldots, T-1$, $\breve{\alpha}_t(j)$ is given by

$$\breve{\alpha}_t(j) = \prod_{l=1}^{t-1} c_l\, \alpha_t(j) = \sum_{d=1}^{t} \widehat{\alpha}^*_{t-d+1}(j)\, p_j(d,\, t-d+1) \breve{u}_t(j,\, d), \tag{28}$$

and for $t = T$, we have

$$\breve{\alpha}_T(j) = \prod_{l=1}^{T-1} c_l\, \alpha_T(j) = \sum_{d=1}^{T} \widehat{\alpha}^*_{T-d+1}(j) \bar{P}_j(d,\, t-d+1) \breve{u}_T(j,\, d). \tag{29}$$

Also

$$\widehat{\alpha}^*_{t+1}(j) = \sum_{i=1}^{m} \widehat{\alpha}_t(i) a_{ij}. \tag{30}$$

Note that the recursion formula for the variable $c_t^*(j)$ uses the same arguments leading to the forward recursion in (9) and (10), calculating the survival function of state duration probability distributions at each time t, instead. Computations of the scaling factor c_t and variable $c_t^*(j)$ can be performed while doing the forward recursion because they do not require an extra loop in the algorithm. Also, their computations do not change the order of the forward algorithm with the

same computational complexity of the forward algorithm proposed by Guédon (2003). However, as compared to the forward algorithm of Guédon (2003), we introduced two new variables $c_t^*(j)$ and $u_t(j, d)$ for the ease of expressions and computations.

Note that the above scaling procedure satisfies the relation $\sum_{j=1}^{m} \Pr(S_t = j \mid X_{1:t}, \Theta) = \sum_{j=1}^{m} c_t c_j^*(t) = 1$ at each time step t. It does not make the condition that $\sum_{j=1}^{m} \widehat{\alpha}_t(j) = 1$ for $t = 1, \ldots, T-1$ because of the boundry condition that state j must end at time t. However, we have $\widehat{\alpha}_T(j) = 1$ for the last visited state not finishing at time $t = T$. It is important to note that

$$c_t = 1 \Big/ \sum_{j=1}^{m} c_t^*(j) \,,$$

$$c_t = 1 \Big/ \Pr(X_{1:t} \mid \Theta) \prod_{l=1}^{t-1} c_l \,,$$

$$\prod_{l=1}^{t} c_l = 1 / \Pr(X_{1:t} \mid \Theta) \,, \tag{31}$$

and

$$\prod_{l=1}^{t-1} c_l = 1 / \Pr\left(X_{1:(t-1)} \mid \Theta\right) . \tag{32}$$

Dividing (31) by (32), we have

$$c_t^{-1} = \Pr\left(X_t \mid X_{1:(t-1)}\right).$$

Whence, the scaling factor c_t at time t defines the likelihood of an observation X_t conditionally on the past observations $X_{1:(t-1)}$ as is the case for HMMs. This scaling procedure transforms $\alpha_t(j)$ from a joint probability to a conditional probability (Li and Yu, 2015). That is,

$$\begin{aligned}
\widehat{\alpha}_t(j) &= \prod_{l=1}^{t} c_l \alpha_t(j) \\
&= \frac{\Pr(X_{1:t},\, S_t = j,\, S_{t+1} \neq j \mid \Theta)}{\Pr(X_{1:t} \mid \Theta)} \\
&= \Pr(S_t = j,\, S_{t+1} \neq j \mid X_{1:t},\, \Theta) \,,
\end{aligned} \tag{33}$$

which is equivalent to the forward variable proposed by Guédon (2003). Using the scaling factors, the log-likelihood function can be obtained as we do in HMMs. Thus, we have

$$\log \mathcal{L}(\Theta) = -\sum_{t=1}^{T} \log c_t. \tag{34}$$

The pseudo code for the forward recursion along with the calculation of the log-likelihood function is outlined in Algorithm 1.

The backward variable can be scaled by the scaling factor c_t computed for the forward variable.

From (11), the scaled backward variable $\widehat{\beta}_t(j)$ for $t = T-1, \ldots, 1$ is obtained as

$$\begin{aligned}
\widehat{\beta}_t(j) &= \prod_{l=t+1}^{T} c_l\, \beta_t(j) \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \sum_{d=1}^{T-t-1} a_{ji} p_i(d,\, t+1) \prod_{k=t+1}^{t+d} c_k\, u_{t+d}(i,\, d) \prod_{h=t+d+1}^{T} c_h\, \beta_{t+d}(i) \\
&\qquad + \sum_{\substack{i=1 \\ i\neq j}}^{m} a_{ji} \bar{P}_i(T-t,\, t+1) \prod_{l=t+1}^{T} c_l\, u_T(i,\, T-t) \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \sum_{d=1}^{T-t-1} a_{ji} p_i(d,\, t+1) \widehat{u}_{t+d}(i,\, d) \widehat{\beta}_{t+d}(i) + \sum_{\substack{i=1 \\ i\neq j}}^{m} a_{ji} \bar{P}_i(T-t,\, t+1) \widehat{u}_T(i,\, T-t),
\end{aligned} \tag{35}$$

where $\widehat{u}_t(j,\, d)$ are the scaled observation probabilities which can be obtained from (24) by the following relation

$$\widehat{u}_t(j,\, d) = c_t\, \breve{u}_t(j,\, d). \tag{36}$$

Also, from (13) and (14) for $t = T-1, \ldots, 1$, we have

$$\widehat{\beta}_t(i) = \sum_{\substack{j=1 \\ i\neq j}}^{m} a_{ij} \widehat{\beta}^{*}_{t+1}(j), \tag{37}$$

and

$$\widehat{\beta}^{*}_{t+1}(j) = \sum_{d=1}^{T-t-1} \widehat{\beta}_{t+d}(j) p_j(d,\, t+1)\, \widehat{u}_{t+d}(j,\, d) + \bar{P}_j(T-t,\, t+1)\, \widehat{u}_T(j,\, T-t). \tag{38}$$

Note that the scaled backward variable becomes the ratio of two probabilities and does not seem to have any natural interpretation. That is,

$$\begin{aligned}
\widehat{\beta}_t(j) &= \prod_{l=t+1}^{T} c_l \beta_t(j) \\
&= \frac{\Pr\left(X_{(t+1):T} \,|\, S_t = j,\, S_{t+1} \neq j,\, \Theta\right)}{\Pr\left(X_{(t+1):T} \,|\, X_{1:t},\, \Theta\right)}.
\end{aligned} \tag{39}$$

Also, note that once the scaled observation probabilities and their products are stored during the forward recursion in (24), we can use them directly for the backward recursion instead of computing them again. The probability $\gamma_t(j)$ in (19) can be computed using the scaled forward and backward variables during the backward recursion. When we use the scaled forward variable, we have

$$\mathcal{L}(\Theta) = \sum_{j=1}^{m} \widehat{\alpha}_T(j) = 1. \tag{40}$$

Thus, the recursive formula for $\gamma_t(j)$ in (19) reduces to

$$\gamma_t(j) = \gamma_{t+1}(j) + \widehat{\alpha}_t(j)\widehat{\beta}_t(j) - \widehat{\alpha}^*_{t+1}(j)\widehat{\beta}^*_{t+1}(j), \tag{41}$$

with the initial condition

$$\gamma_T(j) = \widehat{\alpha}_T(j). \tag{42}$$

The pseudo code for the backward recursion along with the computation of $\gamma_t(j)$ is outlined in Algorithm 2.

Algorithm 1 The forward algorithm for an $m-$state IHSMM

for $j = 1$ **to** m **do**
  $\widehat{\alpha}^*_1(j) = \pi_j$ (22)
end for
for $t = 1$ **to** T **do**
  for $j = 1$ **to** m **do**
    $\breve{\alpha}_t(j) = 0$, $c^*_t(j) = 0$ and $\widehat{\alpha}^*_{t+1}(j) = 0$
    if $t < T$ **then**
      for $d = 1$ **to** t **do**
        if $d = 1$ **then**
          $\breve{u}_t(j,\, d) = f_j(x_t)$ (25)
        else
          $\breve{u}_t(j,\, d) = c_{t-1}\,\breve{u}_{t-1}(j,\, d-1) f_j(x_t)$ (24)
          $\breve{\alpha}_t(j) = \breve{\alpha}_t(j) + \widehat{\alpha}^*_{t-d+1}(j) p_j(d,\, t-d+1)\breve{u}_t(j,\, d)$ (28)
          $c^*_t(j) = c^*_t(j) + \widehat{\alpha}^*_{t-d+1}(j)\bar{P}_j(d,\, t-d+1)\breve{u}_t(j,\, d)$ (23)
        end if
      end for
    else
      for $d = 1$ **to** t **do**
        if $d = 1$ **then**
          $\breve{u}_T(j,\, d) = f_j(x_T)$ (25)
        else
          $\breve{u}_T(j,\, d) = c_{T-1}\,\breve{u}_{T-1}(j,\, d-1) f_j(x_T)$ (24)
          $\breve{\alpha}_T(j) = \breve{\alpha}_T(j) + \widehat{\alpha}^*_{T-d+1}(j)\bar{P}_j(d,\, T-d+1)\breve{u}_T(j,\, d)$ (29)
          $c^*_T(j) = c^*_T(j) + \widehat{\alpha}^*_{T-d+1}(j)\bar{P}_j(d,\, T-d+1)\breve{u}_T(j,\, d)$ (23)
        end if
      end for
    end if
  end for
  $c_t = 1 \Big/ \sum_{j=1}^m c^*_t(j)$ (26)
  for $j = 1$ **to** m **do**
    $\widehat{\alpha}_t(j) = c_t \breve{\alpha}_t(j)$ (27)
    for $i = 1$ **to** m **do**
      if $(t < T)$ **then**
        $\widehat{\alpha}^*_{t+1}(j) = \widehat{\alpha}^*_{t+1}(j) + \widehat{\alpha}_t(i)\, a_{ij}$ (30)
      end if
    end for
  end for
end for
The log-likelihood is $\log \mathcal{L}(\Theta) = -\sum_{t=1}^T \log(c_t)$. (34)

Algorithm 2 The backward algorithm and calculation of $\gamma_t(j)$ for an $m-$ state IHSMM

for $t = T$, $j = 1$ **to** m **do**
  $\gamma_T(j) = \widehat{\alpha}_T(j)$ (42)
end for
for $t = T-1$ **to** 1 **do**
  $\widehat{\beta}^*_{t+1}(j) = 0$
  for $j = 1$ **to** m **do**
    for $d = 1$ **to** $T-t$ **do**
      if $(d < T-t)$ **then**
        $\widehat{u}_{t+d}(j,\, d) = c_{t+d}\, \breve{u}_{t+d}(j,\, d)$ (36)
        $\widehat{\beta}^*_{t+1}(j) = \widehat{\beta}^*_{t+1}(j) + \widehat{\beta}_{t+d}(j)\, p_j(d,\, t+1)\, \widehat{u}_{t+d}(j,\, d)$ (38)
      else
        $\widehat{u}_T(j,\, T-t) = c_T\, \breve{u}_T(j,\, T-t)$ (36)
        $\widehat{\beta}^*_{t+1}(j) = \widehat{\beta}^*_{t+1}(j) + \bar{P}_j(T-t,\, t+1)\, \widehat{u}_T(j,\, T-t)$ (38)
      end if
    end for
  end for
  for $j = 1$ **to** m **do**
    $\widehat{\beta}_t(j) = 0$ and $\gamma_t(j) = 0$
    for $i = 1$ **to** m **do**
      $\widehat{\beta}_t(j) = \widehat{\beta}_t(j) + a_{ji}\widehat{\beta}^*_{t+1}(i)$ (37)
      $\gamma_t(j) = \gamma_{t+1}(j) + \widehat{\alpha}_t(j)\widehat{\beta}_t(j)$ - $\widehat{\alpha}^*_{t+1}(j)\widehat{\beta}^*_{t+1}(j)$ (41)
    end for
  end for
end for

S.4 The Viterbi path

The Viterbi algorithm is a dynamic programming algorithm for obtaining the most probable sequence of hidden states, called as Viterbi path, in the context of HMM (Viterbi, 1967). To find the most optimal state sequence $\mathbf{S}^*$ given an observed sequence $\mathbf{X}$ is one of the primary objectives after the estimation of parameters of HMMs, i.e.

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} \Pr\left(\mathbf{S} \,|\, \mathbf{X},\, \Theta\right),$$

where

$$\max_{\mathbf{S}} \Pr\left(\mathbf{S} \,|\, \mathbf{X},\, \Theta\right) = \max_{\mathbf{S}} \frac{\Pr\left(\mathbf{S},\, \mathbf{X} \,|\, \Theta\right)}{\Pr\left(\mathbf{X}\right)}.$$

There exists many variants of the Viterbi path for HSMMs in the literature. Following Guédon (2003) and Yu (2010), we extend the Viterbi algorithm for our proposed IHSMM. To find the optimal state sequence for an $m-$state IHSMM, we define a forward variable representing the maximum likelihood that the partial state sequence ending at time t in state j for duration d by

$$\begin{aligned}
\delta_t(j,d) &= \max_{S_{1:(t-d)}} \Pr\left(S_{1:(t-d)},\, S_{(t-d+1):t} = j,\, S_{t+1} \neq j,\, X_{1:t} \,|\, \Theta\right) \\
&= \max_{i \neq j} \max_{h} \max_{S_{1:(t-d-h)}} \Pr\left(S_{1:(t-d-h)},\, S_{(t-d-h+1):(t-d)} = i,\, S_{(t-d+1):t} = j,\, S_{t+1} \neq j,\, X_{1:t} \,|\, \Theta\right) \\
&= \max_{i \neq j} \max_{h} \left\{\delta_{t-d}(i,\, h) a_{ij} p_j(d, t-d+1) f_j(x_{(t-d+1):t})\right\}, \qquad (43)
\end{aligned}$$

for $2 \leqslant t \leqslant T-1$, $j = 1, \ldots, m$ and $d = 1, \ldots, t-1$, with the initial condition given by

$$\delta_t(j,\, d) = \pi_j p_j(d,\, 1) f_j(x_{(t-d+1):t}), \tag{44}$$

for $t = d = 1, \ldots, T-1$ and $j = 1, \ldots, m$. And for right censored sojourn time in the last visited state at $t = T$, we have

$$\begin{aligned}
\delta_T(j, d) &= \max_{S_{1:(T-d)}} \Pr\left(S_{1:(T-d)},\, S_{(T-d+1):T} = j,\, X_{1:T} \,|\, \Theta\right) \\
&= \max_{i \neq j} \max_{h} \max_{S_{1:(T-d-h)}} \Pr\left(S_{1:(T-d-h)},\, S_{(T-d-h+1):(T-d)} = i,\, S_{(T-d+1):T} = j,\, X_{1:T} \,|\, \Theta\right) \\
&= \max_{i \neq j} \max_{h} \left\{\delta_{T-d}(i,\, h) a_{ij}\, \bar{P}_j(d, T-d+1) f_j(x_{(T-d+1):T})\right\},
\end{aligned} \tag{45}$$

with the initial condition

$$\delta_T(j,\, d) = \pi_j\, \bar{P}_j(T,\, 1) f_j(x_{(T-d+1):T}), \tag{46}$$

for $d = 1, \ldots, T$. We record the previous state i^* and its sojourn h^* selected by $\delta_{t-d}(j,\, d)$ in the following array

$$\psi(t,\, j,\, d) = (t-d,\, i^*,\, h^*), \tag{47}$$

where $(t-d)$ is the ending time of most probable state i^* having duration h^* and

$$(i^*,\, h^*) = \arg\max_{i \neq j} \max_{h} \left\{\delta_{t-d}(i,\, h) a_{ij} p_j(d,\, t-d+1) f_j(x_{(t-d+1):t})\right\}. \tag{48}$$

Finally, overall global optimal state sequence probabilities are in $\delta_T(j,\, d)$, $j \geqslant 1$ and $d \geqslant 1$, which can be traced back as for $t_0^* = T$

$$(j_0^*,\, d_1^*) = \arg\max_{j} \max_{d} \delta_T(j,\, d), \tag{49}$$

and for $t_1^*, \ldots, t_n^*$

$$\begin{aligned}
(t_1^*,\, j_1^*,\, d_2^*) &= \psi(t_0^*,\, j_0^*,\, d_1^*) \\
&\vdots \\
(t_{n(T)}^*,\, j_{n(T)}^*,\, d_{n(T)+1}^*) &= \psi(t_{n(T)-1}^*,\, j_{n(T)-1}^*,\, d_{n(T)}^*),
\end{aligned} \tag{50}$$

where $S_1 = j_{n(T)}^*$ is the first visited state and $S_{n(T)} = j_0^*$ is the last visited state. Thus, $(j_{n(T)}^*,\, d_{n(T)+1}^*) \ldots (j_0^*,\, d_1^*)$ is the most likely occurred state sequence given observed data for an IHSMM. Note that d_1^* is the minimum time spent in the last visited state. We outline the above procedure to find the Viterbi path for an IHSMM in the following Algorithm 3. In order to avoid the underflow problem of multiplied probabilities, the logarithm of the probabilities can be used.

S.5 Supplement to data analysis

In the main manuscript, we analyzed a global volcanic eruption catalogue by fitting different types of HMMs. The supporting tables and figures are provided in this Supplementary file. The number of parameters (k), maximum log-likelihood (MLL) and Akaike Information Criterion

Algorithm 3 The Viterbi path algorithm for an $m-$state IHSMM

for $j = 1$ **to** m and $t = d = 1$ **to** T **do**
  if $t < T$ **then**
    $\delta_t(j,\, d) = \pi_j\, p_j(d,\, 1) f_j(x_{(t-d+1):t})$ (44)
  else
    $\delta_T(j,\, d) = \pi_j\, \bar{P}_j(T,\, 1) f_j(x_{(T-d+1):T})$ (46)
  end if
end for
for $t = 2$ **to** T **do**
  if $t < T$ **then**
    for $j = 1$ **to** m **do**
      for $d = 1$ **to** $t - 1$ **do**
        $\delta_t(j,\, d) = \max_{i \neq j} \max_h \left\{\delta_{t-d}(i,\, h) a_{ij}\, p_j(d,\, t-d+1) f_j(x_{(t-d+1):t})\right\}$ (43)
        $(i^*,\, h^*) = \arg\max_{i \neq j} \max_h \left\{\delta_{t-d}(i,\, h) a_{ij}\, p_j(d, t-d+1) f_j(x_{(t-d+1):t})\right\}$ (48)
        $\psi(t,\, j,\, d) = (t-d,\, i^*,\, h^*)$ (47)
      end for
    end for
  else
    for $j = 1$ **to** m **do**
      for $d = 1$ **to** $t - 1$ **do**
        $\delta_T(j,\, d) = \max_{i \neq j} \max_h \left\{\delta_{T-d}(i,\, h) a_{ij}\, \bar{P}_j(d,\, T-d+1) f_j(x_{(T-d+1):T})\right\}$ (45)
        $(i^*,\, h^*) = \arg\max_{i \neq j} \max_h \left\{\delta_{T-d}(i,\, h) a_{ij}\, \bar{P}_j(d, T-d+1) f_j(x_{(T-d+1):T})\right\}$ (48)
        $\psi(T,\, j,\, d) = (T-d,\, i^*,\, h^*)$ (47)
      end for
    end for
  end if
end for
Trace back letting $t_0^* = T$, $\quad (j_0^*,\, d_1^*) = \arg\max_j \max_d \delta_T(j,\, d)$. (49)
for $n(T) = 1, 2, 3, \ldots$ **do**
  $(t^*_{n(T)},\, j^*_{n(T)},\, d^*_{n(T)+1}) = \psi(t^*_{n(T)-1},\, j^*_{n(T)-1},\, d^*_{n(T)})$ (50)
  continue the tracing back until $t^*_{n(T)} - d^*_{n(T)+1} < 1$.
end for
The required most probable state sequence for an $m-$state IHSMM is $(j^*_{n(T)},\, d^*_{n(T)+1}) \ldots (j_0^*,\, d_1^*)$.

(AIC) values for the fitted models in Case I in Section 8.2 of the main manuscript are listed in Table S.1. The Viterbi paths for the 4−state and 5−state IHSMMs in Case I suggested by AIC are shown in Figure S.1. The residual analyses for each of the fitted HMMs, HSMMs and IHSMMs in case I are provided in Figure S.2. The results in these figures have been discussed in Section 8.2 of the main manuscript.

Looking at Cases II and IV individually in Table 4 of the main manuscript, we observe that in Case II the 4−state IHSMM appears to be the best fitted model with the smallest AIC value along with the 5−state IHSMM as another possible suitable model for the data based on the AIC difference being less than 2. In Case IV, the 5−state and 6−state IHSMMs have AIC values 0.20 apart and can be considered as suitable models for further selection. However, Case III provides the 4−state IHSMM with the smallest AIC as the best fit model.

The residual analysis for each of the fitted models in all cases in Table 4 of the main manuscript are given in Figures S.3 to S.5. From these figures, we observe that for the IHSMMs selected by AIC except for the 4−state IHSMM in Case II, the residual processes seem to be well approximated by a stationary Poisson process with unit rate. The residual process for other higher state IHSMMs

Table S.1: No. of parameters (k), MLL and AIC in Case I

	HMM			HSMM			IHSMM		
Model	k	MLL	AIC	k	MLL	AIC	k	MLL	AIC
3−state	9	-791.41	1600.83	9	-799.37	1616.73	18	-769.94	1575.88
4−state	16	-777.06	1586.11	26	-790.14	1612.29	28	-750.74	1557.48
5−state	25	-764.11	1578.22	25	-777.83	1605.65	40	-738.41	1556.81
6−state	36	-751.51	1575.03	36	-767.56	1607.12	54	-728.65	1565.30
7−state	49	-741.76	1581.53	49	-758.35	1614.71	70	-723.50	1587.00

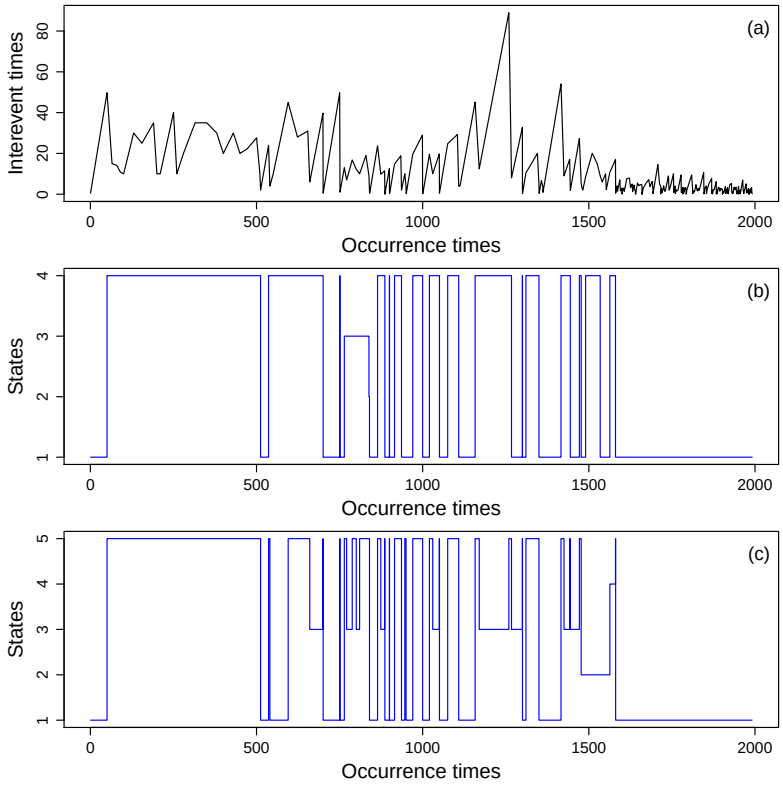

Figure S.1: (a) The observed interevent times, (b) The Viterbi path for the 4−state IHSMM and (c) The Viterbi path for the 5−state IHSMM in Case 1.

in these cases do not improve by much. Also, the residual process for HMMs and HSMMs do not appear to be well approximated by a stationary Poisson process in all cases. Since none of these HMMs and HSMMs have AIC values close to the smallest AIC value in each case individually and in all cases collectively, we do not consider these models for further analysis.

Based on AIC values and Figures S.3 to S.5, we consider the 4−state IHSMM in Case III and the 5−state IHSMMs in Cases II and IV. Since the 4−state IHSMM has been selected and discussed in Sections 8.2 and 8.3, respectively, of the main manuscript, we check the further assumptions of a stationary Poisson process for the 5−state IHSMMs in Cases II and IV in this supplementary file. Using the KS test of uniformity, the empirical distributions of U_k from these two models are plotted in Figures S.6 and S.7, which shows uniformity of U_k. Hence, there is no evidence to assert that the transformed interevent times, E_k, are not exponentially distributed. Also, in Figures S.6 and S.7, the scatter plots of E_{k+1} against E_k and U_{k+1} against U_k for the two 5−state models show no particular pattern of points for any association, suggesting the independence of E_k from

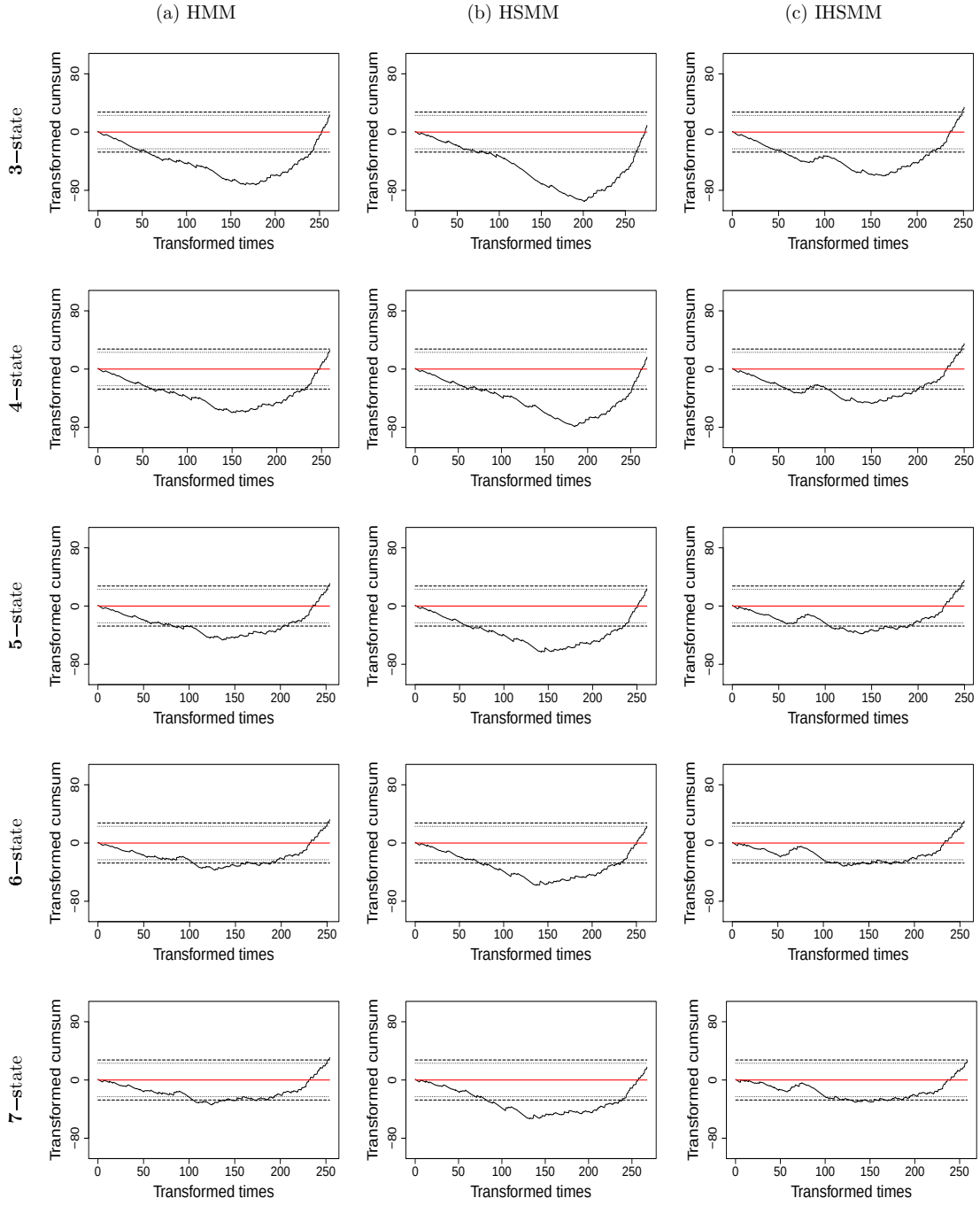

Figure S.2: Case I: The deviated cumulative number of events in the residual process from the stationary process versus the transformed times for HMMs, HSMMs and IHSMMs with 3, 4, 5, 6 and 7 hidden states fitted to the global volcanic eruption catalogue. The central line at zero is the theoretical curve under the null hypothesis of stationary process. The dotted and dashed lines represent the two-sided 95% and 99% confidence limits of the Kolmogrov-Smirnov (KS) statistic, respectively.

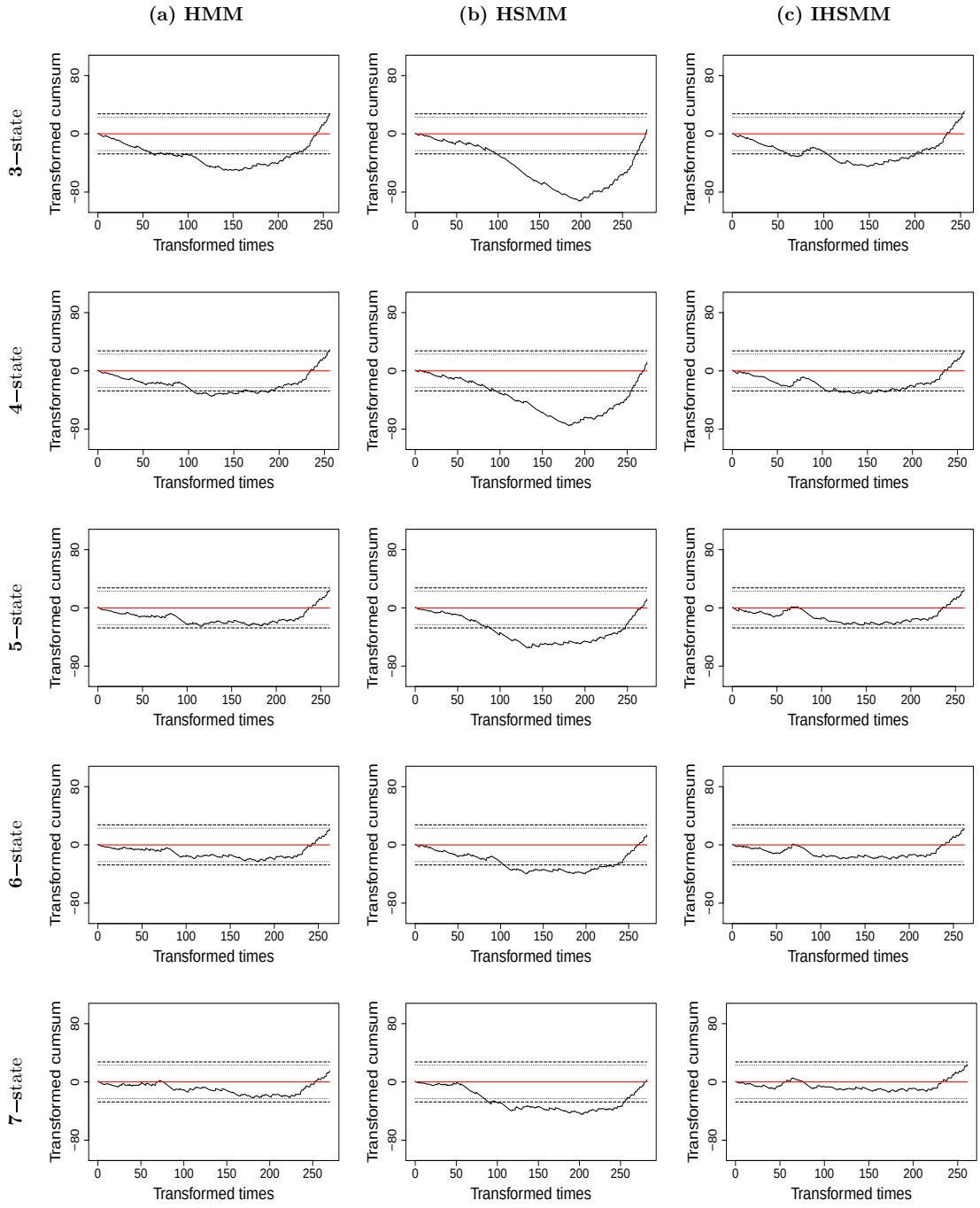

Figure S.3: Case II: The deviated cumulative number of events in the residual process from the stationary process versus the transformed times for HMMs, HSMMs and IHSMMs with 3, 4, 5, 6 and 7 hidden states fitted to the global volcanic eruption catalogue. The central line at zero is the theoretical curve under the null hypothesis of stationary process. The dotted and dashed lines represent the two-sided 95% and 99% confidence limits of the KS statistic, respectively.

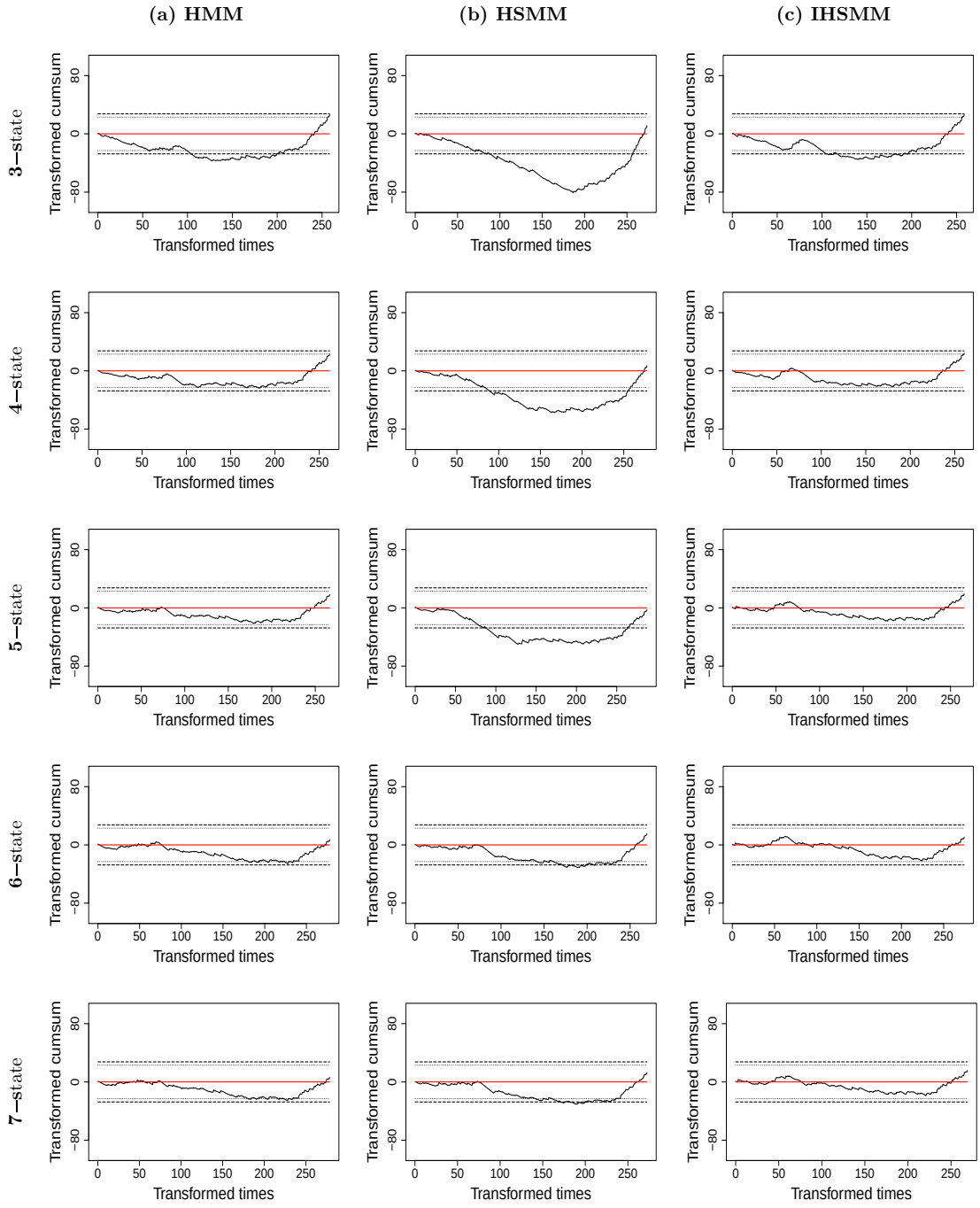

Figure S.4: Case III: The deviated cumulative number of events in the residual process from the stationary process versus the transformed times for HMMs, HSMMs and IHSMMs with 3, 4, 5, 6 and 7 hidden states fitted to the global volcanic eruption catalogue. The central line at zero is the theoretical curve under the null hypothesis of stationary process. The dotted and dashed lines represent the two-sided 95% and 99% confidence limits of the KS statistic, respectively.

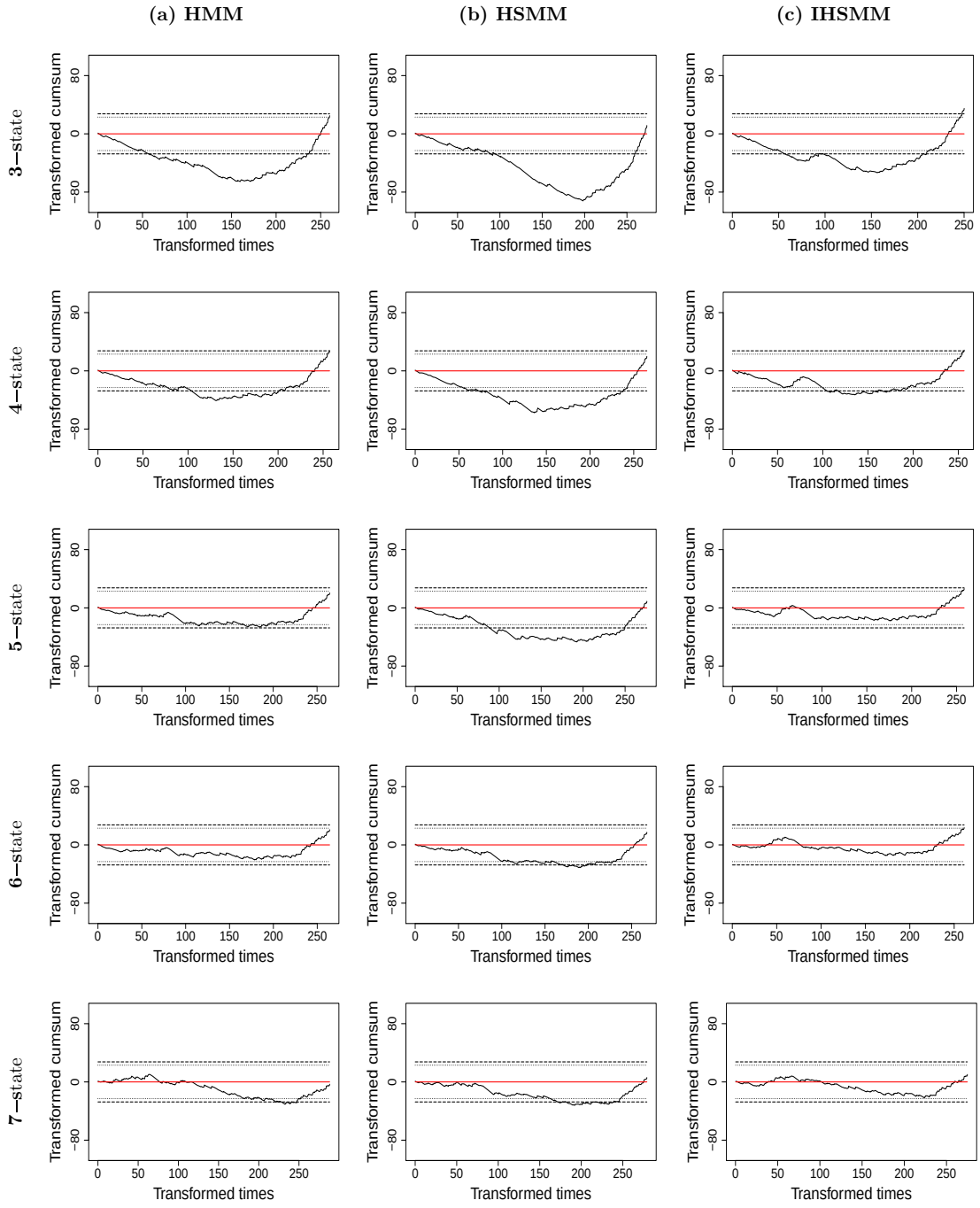

Figure S.5: Case IV: The deviated cumulative number of events in the residual process from the stationary process versus the transformed times for HMMs, HSMMs and IHSMMs with 3, 4, 5, 6 and 7 hidden states fitted to the global volcanic eruption catalogue. The central line at zero is the theoretical curve under the null hypothesis of stationary process. The dotted and dashed lines represent the two-sided 95% and 99% confidence limits of the KS statistic, respectively.

the two 5−state models in Cases II and IV. The t-test for the null hypothesis of zero correlation between E_{k+1} and E_k in these models produces $P-$values of 0.421 and 0.533, further confirming that there is no evidence to reject the hypothesis that E_k are independent. We conclude that the residual processes for the the two 5−state IHSMMs in Cases II and IV follow a stationary Poisson process with unit rate satisfying the assumptions of independence and exponentiality.

We note that the 5−state IHSMMs in Cases II and IV represent a maximum number of missing events up to 8 and 10 between a pair of consecutively observed events in the record, respectively and have 40 parameters. The selected 4−state IHSMM in Case III models a maximum number of missing events up to 9 (the average of the maximum number of missing events of 8 and 10 in the above 5−state models) between a pair of consecutively observed events with 28 parameters. Also, the 4−state IHSMM has the lowest AIC value among all models in all cases. Thus, the overall analysis suggested the 4−state IHSMM in Case III as the best approximation of the given global volcanic eruption record in terms of the number of parameters, AIC, residual analysis and the number of missing events represented by each state.

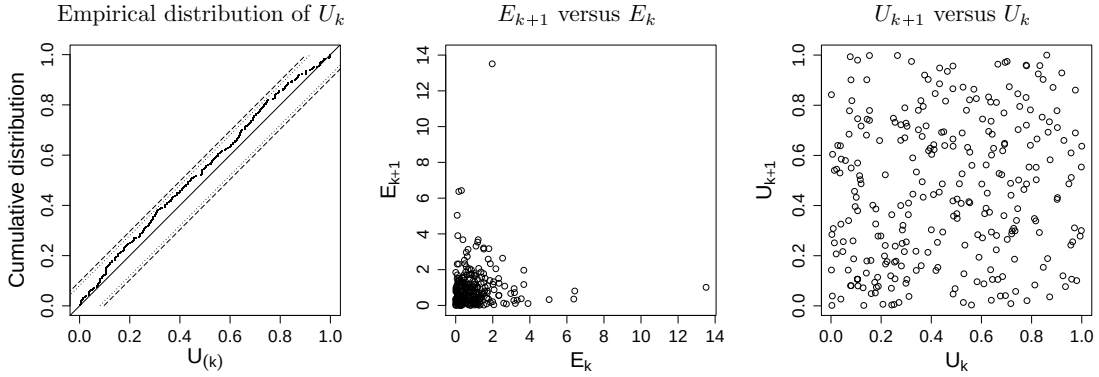

Figure S.6: Residual check for the 5−state IHSMM in Case II. Left: Empirical distribution of U_k, with the dotted and dashed lines indicating 95% and 99% confidence intervals of the KS statistic, assuming uniform distribution. Middle: Scatter plot of E_{k+1} versus E_k. Right: Scatter plot of U_{k+1} versus U_k.

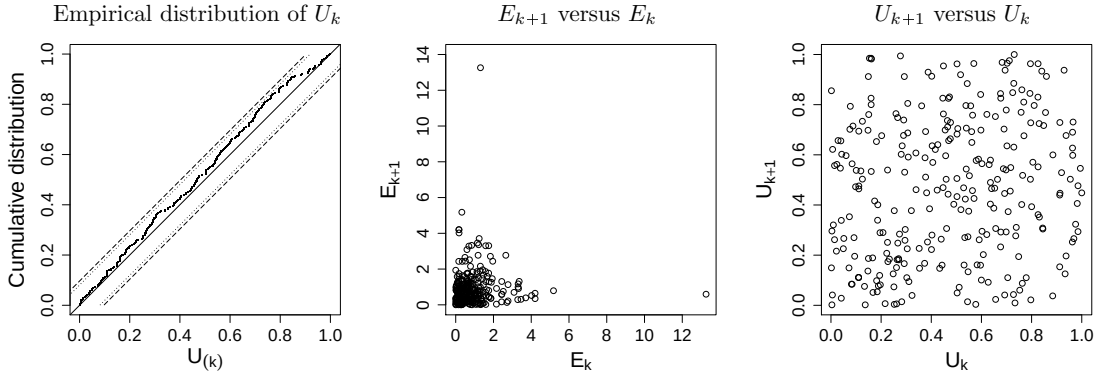

Figure S.7: Residual check for the 5−state IHSMM in Case IV. Left: Empirical distribution of U_k, with the dotted and dashed lines indicating 95% and 99% confidence intervals of the KS statistic, assuming uniform distribution. Middle: Scatter plot of E_{k+1} versus E_k. Right: Scatter plot of U_{k+1} versus U_k.

References

Baum, L. E., Pretrie, T., Soules, G., Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chain. *Annals of Mathematical Statistics*, 41(1), 164–177. doi:10.1214/aoms/1177697196

Bulla, J. (2006) *Application of hidden Markov models and hidden semi-Markov models to financial time series*. PhD-Thesis, Georg-August-Universitat Gottingen.`https://mpra.ub.uni-muenchen.de/id/eprint/7675`

Ferguson, J. D. (1980). Variable duration models for speech. *Proceedings: Symposium on the Application of Hidden Markov Models to Text and Speech*, pp 143–179. New Jersey: Princeton.

Filardo, A. J. (1994). Business-cycle phases and their transitional dynamics. *Journal of Business and Economics Statistics*, 12(3), 299–308. doi:10.2307/1392086

Furlan, C. (2010). Extreme value methods for modelling historical series of large volcanic magnitudes. *Statistical Modelling*, 10(2), 113–132. doi:10.1177/1471082X0801000201

Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3), 604–639. doi:10.1198/1061860032030

Levinson, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1(1), 29–45. doi:10.1016/S0885-2308(86)80009-2

Li. B. C., Yu, S. Z. (2015). A robust scaling approach for implementation of HsMMs. *IEEE Signal Processing Letters*, 22(9), 1264–1268. doi:10.1109/LSP.2015.2397278

Mitchell, C. D., Jamieson, L. H. (1993). Modeling duration in a hidden Markov model with the exponential family. *Proceedings: IEEE International Conference on Acoustics, Speech and Signal Processing*, pp 331–334. doi:10.1109/ICASSP.1993.319304

Murphy, K. P. (2002). Hidden semi-Markov Models (HSMMs). `https://www.cs.ubc.ca/~murphyk/papers/segment.pdf`

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13, 260–269. doi:10.1109/tit.1967.1054010

Yu, S. Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174, 215–243. doi:10.1016/j.artint.2009.11.011