Bayesian Posterior Mean Estimates for Poisson Hidden Markov Models

(The final version has been published in *Computational Statistics & Data Analysis*, vol. 53, no. 4, pp. 941-955, Feb. 2009.
doi:10.1016/j.csda.2008.11.012)

Junko Murakami

Abstract

This paper focuses on the Bayesian posterior mean estimates (or Bayes' estimate) of the parameter set of Poisson hidden Markov models in which the observation sequence is generated by a Poisson distribution whose parameter depends on the underlining discrete-time time-homogeneous Markov chain. Although the most commonly used procedures for obtaining parameter estimates for hidden Markov models are versions of the expectation maximization and Markov chain Monte Carlo approaches, this paper exhibits an algorithm for calculating the exact posterior mean estimates which, although still cumbersome, has polynomial rather than exponential complexity, and is a feasible alternative for use with small scale models and data sets. This paper also shows simulation results, comparing the posterior mean estimates obtained by this algorithm and the maximum likelihood estimates obtained by expectation maximization approach.

Key words: Poisson hidden Markov models, algorithms, Bayesian posterior mean estimates, Bayes' estimates, parameter estimation

1 Introduction

Hidden Markov models (HMMs) are stochastic models in which an underlying Markov chain, which is 'hidden', emits an output sequence that can be observed. The models come in various forms (Ephraim & Merhav, 2002) and are widely used in diverse areas; for example, traffic modeling, event detection, inventory control, and precipitation modeling, to name a few. Poisson hidden Markov models (PHMMs) are one incident of HMMs in which the emission is

Email address: junko.murak@gmail.com (Junko Murakami).

governed by Poisson distributions (MacDonald & Zucchini, 1997; Cooper & Lipsitch, 2004; Albert, 1991). Various applications using this model exist; for example, traffic modeling (Heffes & Lucantoni, 1986; Scott & Smyth, 2003), event detection (Jana & Dey, 2000), and inventory control (Ching, 1997).

The main focus of this paper is on the Bayesian posterior mean estimates (BPM) as the parameter estimate of PHMM. However, we start with more general type of HMMs in which the Markov chain is discrete-time and time-homogeneous, and the observation sequence has probability distributions defined by a function of the observed value. The parameter value of this observation function depends on the current state. The algorithm described here applies to any of such HMMs. Then, a specific case, PHMM, will be investigated.

The algorithm described in this paper, which is a result of a continuation of the study done on the HMMs with a time-homogeneous emission matrix (Murakami & Taylor, 2006), significantly reduces the computational complexity in finding the exact BPM; i.e., from exponential (with a naive approach) to polynomial in the data size, say n. Also, it is a recursive algorithm, which does not require a full computation to be repeated when an additional observation is made. Hence, the algorithm makes it possible for us to efficiently compare and discuss the two different approaches for the parameter estimation, the BPM and maximum likelihood estimate (MLE).

However, this algorithm does have significant limitations. It is limited to situations where both state space and observations are discrete. The computational complexity, although reduced to polynomial complexity in the number of observations, is still of exponential complexity in the number of states and in the largest observed value (see the end of Section 3.4 and also Section 5.2 for details), and in practice the technique is only feasible for small scale problems. Hence, it is not feasible for many types of HMM applications for which various methods for the MLE, such as expectation maximization (EM) algorithms (Dempster et al., 1977; Baum et al., 1970; Baum, 1972; Rabiner, 1989) or approximation methods for the BPM such as Markov chain Monte Carlo (MCMC) methods (Scott, 2002; Chib, 1996; Robert & Titterington, 1998) are feasible.

Still, it is interesting to compare the characteristics of MLE and BPM when it is feasible. Taking advantage of the significant reduction in the computational complexity, simulations are implemented using the MLE obtained by EM algorithm and the exact BPM for two-state PHMMs with small n and observation space, and the results show that the stability problem of MLE that is often found for the applications with a small n (closely related to the well-known 'overfitting' problem) does not exist in case for the BPM, as expected. The simulation results also show some advantages of the BPM especially when the state-dependent parameters for the observation functions are close to each other.

In addition, in another point of view, this algorithm can be considered as a proof that the computational complexity in n for finding the exact BPM can be reduced from exponential to polynomial.

As for the flow of this paper, the HMMs studied in this paper are first described in details in Section 2, and the method, including the algorithm, used to obtain the BPM is described in Section 3. Finally, we focus on the PHMMs in Section 4, where the formulas for this specific HMMs are given. Various simulation results and discussions on the results follow in Section 5, where comparisons are made between estimates obtained by the algorithm for the BPM and standard EM procedures. Finally, conclusions are presented in Section 6.

2 Hidden Markov Models

Let $X^{1:n} = (x_1, x_2, x_3, \ldots, x_n)$, $x_i \in \zeta_X$, and $Y^{1:n} = (y_1, y_2, y_3, \ldots, y_n)$, $y_u \in \zeta_Y$, be the Markov chain and observation sequences of length n, respectively, where $\zeta_X = \{0, 1, 2, \ldots, \nu - 1\}$ and $\zeta_Y \subseteq \{0, 1, 2, \ldots, \infty\}$ for some positive integer ν.

Consider HMMs, which have the following properties for any positive integer t:

$$P\left(x_{t+1} \mid Y^{1:t}, X^{1:t}\right) = P\left(x_{t+1} \mid x_t\right) \quad \text{and}$$
$$P\left(y_t \mid Y^{1:t-1}, Y^{t+1:n}, X^{1:n}\right) = P\left(y_t \mid x_t\right).$$

The Markov chain considered in this paper is governed by a $\nu \times \nu$ time-homogeneous transition matrix $A = \{a_{ij}\}$, where $\nu > 1$ is any positive integer and a_{ij} are probability constants such that

$$a_{ij} = P(x_{t+1} = j \mid x_t = i) \quad \text{for } i, j \in \zeta_X.$$

As for the probability distribution for the observation sequence, let

$$f_{i,\lambda_i}(u) = P(y_t = u \mid x_t = i) \quad \text{for } i \in \zeta_X, u \in \zeta_Y,$$

where $\lambda_i \in \mathbb{R}$ is the parameter for the probability distribution function when the current state is i. Let $\Lambda = \{\lambda_i\}_{i \in \zeta_X}$. Finally, as for the initial state distribution, let Π defined as $\Pi = \{\pi_i\}$ where $\pi_i = P(x_1 = i)$, $i \in \zeta_X$. Denote the set of parameters as $\theta = \{A, \Lambda, \Pi\}$, subject to the normalizing conditions spelled out in Section 3.

Note, due to the independence in the conditional probabilities of the HMMs, we have

$$P(X^{1:n}, Y^{1:n} \mid \theta) = \pi_{x_1} \left(\prod_{t=1}^{n-1} a_{x_t x_{t+1}} \right) \left(\prod_{t=1}^{n} f_{x_t, \lambda_{x_t}}(y_t) \right). \tag{1}$$

3 Bayesian Posterior Mean Estimates

The BPM of the parameter set θ is the expected value of θ given an observation sequence $Y^{1:n}$. Denote the range of $\theta = \{A, \Lambda, \Pi\}$ as $\Theta = \{\mathcal{R}_A, \mathcal{R}_\Lambda, \mathcal{R}_\Pi\}$ so that A is a $\nu \times \nu$ probability matrix under the restriction $a_{ii} \geq a_{i+1,i+1}$ for all $i \in \{0, 1, \ldots, \nu - 2\}$, to avoid 'averaging up' the symmetry (see Section 3.1); Λ is a length-ν vector of real numbers (subject to the restrictions from $f_{i,\lambda_i}(u)$, $i \in \zeta_X$, $u \in \zeta_Y$); and Π is a length-ν probability vector. Using the Bayes' theorem, the expected value of the parameter given an observation sequence can be expressed as

$$\hat{\theta} = \int_{\theta \in \Theta} \theta P\left(\theta \mid Y^{1:n}\right) d\theta = \frac{\int_{\theta \in \Theta} \theta P\left(Y^{1:n} \mid \theta\right) P(\theta) d\theta}{P\left(Y^{1:n}\right)}.$$

Taking the marginal distribution over θ in the denominator, then taking the marginal distribution over $X^{1:n}$ in both the numerator and the denominator, we get

$$\hat{\theta} = \frac{\sum_{X^{1:n} \in \Omega_n} \int_{\theta \in \Theta} \theta \, P\left(X^{1:n}, Y^{1:n} \mid \theta\right) P(\theta) \, d\theta}{\sum_{X^{1:n} \in \Omega_n} \int_{\theta \in \Theta} P\left(X^{1:n}, Y^{1:n} \mid \theta\right) P(\theta) \, d\theta}, \tag{2}$$

where Ω_t is defined to be the set of all possible values of $X^{1:t}$ for any positive integer t. Note that this equation involves a separate integral for each parameter in the set θ; i.e., in the numerator, the integral is actually a set of integrals, while the denominator gives a real number $P\left(Y^{1:n}\right)$.

In order to rewrite $P(X^{1:n}, Y^{1:n} \mid \theta)$ in more details, define $K^{(t)} = \{k_{ij}^{(t)}\}$, $i, j \in \zeta_X$, and $L^{(t)} = \{l_{iu}^{(t)}\}$, $i \in \zeta_X$, $u \in \zeta_Y$, for a positive integer t as follows:

$$k_{ij}^{(t)} = k_{ij}\left(X^{1:t}\right) = \sum_{s=2}^{t} \mathbf{I}\{x_{s-1} = i, x_s = j\} \quad \text{and} \tag{3}$$

$$l_{iu}^{(t)} = l_{iu}\left(X^{1:t}, Y^{1:t}\right) = \sum_{s=1}^{t} \mathbf{I}\{x_s = i, y_s = u\}. \tag{4}$$

Then (1) can also be written as

$$P(X^{1:n}, Y^{1:n} \mid \theta) = \pi_{x_1} \left(\prod_{i,j \,\in\, \zeta_X} a_{ij}^{k_{ij}\left(X^{1:n}\right)} \right) \left[\prod_{\substack{i \,\in\, \zeta_X \\ u \,\in\, \zeta_Y}} \left(f_{i,\lambda_i}(u)\right)^{l_{iu}\left(X^{1:n},\, Y^{1:n}\right)} \right]. \quad (5)$$

It is clear now to see that $K^{(n)}$ and $L^{(n)}$ are complete data sufficient statistics, which the algorithm seeks.

3.1 Identifiability of HMMs

As for identifiability problem of HMMs, in this paper we only consider so-called 'label-switching' problem, which becomes an issue especially when the estimate is the BPM. It is due to the symmetries in the probability distribution space that exist because any permutation of the labels of the states would not change the likelihood of the observation sequence when the prior is set to be symmetric (i.e., invariant under the permutation). Many discussions have been made and various methods have been proposed, usually through the context of HMMs as an extension of mixture models (Celeux et al., 2000; Stephens, 2000; Capppé et al., 2005). However, it is beyond the scope of this paper. In this paper we order the diagonal elements of the transition matrix, which alters the otherwise symmetric prior (see Appendix A). Although other methods have been proposed, this works satisfactorily in our applications.

3.2 Base formulas for BPM

In this section the formulas for the BPM that is used for the simulations are described, mainly to establish the notations. For the sake of simpler expressions, we let $\pi_i = \frac{1}{\nu}$ for $i \in \zeta_X$, though it is not hard to extend the estimator also to find the estimate of Π. Assuming independent and uniform prior distributions as for the parameters for the initial state, from (5), we see that the integration appearing in the denominator of (2) is in the form

$$\begin{aligned} &\int_{\theta\in\Theta} P\left(X^{1:n},\ Y^{1:n} \mid \theta\right) P(\theta) d\theta \\ &\qquad\qquad = \int_{\theta\in\Theta} P\left(X^{1:n} \mid A\right)\, P\left(Y^{1:n} \mid X^{1:n},\, \Lambda\right) P(A,\Lambda)\, d\theta \\ &\qquad\qquad = \frac{1}{\nu}\, \Phi\left(K^{(n)}\right) \cdot \Psi\left(L^{(n)}\right), \qquad (6) \end{aligned}$$

where $\frac{1}{\nu}\, \Phi\left(K^{(n)}\right)$ is the unconditional distribution of the transition data, obtained by integrating over the prior distribution for the transition probabilities.

Similarly $\Psi\left(L^{(n)}\right)$ is the conditional distribution of the observation data, obtained by integrating over the prior distribution assuming the state sequence is known. In other words,

$$\frac{1}{\nu}\,\Phi\left(K^{(n)}\right) = \int_{A\in\mathcal{R}_A} P\left(X^{1:n} \mid A\right) P(A)\,dA \quad \text{and} \tag{7}$$

$$\Psi\left(L^{(n)}\right) = \int_{\Lambda\in\mathcal{R}_\Lambda} P\left(Y^{1:n} \mid X^{1:n}, \Lambda\right) P(\Lambda)\,d\Lambda. \tag{8}$$

Here the counts $K^{(n)}$ and $L^{(n)}$ depends only on $X^{1:n}$, while the count $L^{(n)}$ depends on both $X^{1:n}$ and $Y^{1:n}$. Similarly, the integration that appears in the numerator in (2) can also be viewed as products of integrals. They are

$$\begin{aligned}
&\int_{\theta\in\Theta} A\,P\left(X^{1:n}, Y^{1:n} \mid \theta\right) P(\theta)\,d\theta \\
&\quad = \int_{A\in\mathcal{R}_A} A\,P\left(X^{1:n} \mid A\right) P(A)\,dA \cdot \int_{\Lambda\in\mathcal{R}_\Lambda} P\left(Y^{1:n} \mid X^{1:n}, \Lambda\right) P(\Lambda)\,d\Lambda \\
&\quad = \frac{1}{\nu}\,\widetilde{\Phi}\left(K^{(n)}\right)\Psi\left(L^{(n)}\right),
\end{aligned}$$

for the transition matrix A and

$$\begin{aligned}
&\int_{\theta\in\Theta} \Lambda\,P\left(X^{1:n}, Y^{1:n} \mid \theta\right) P(\theta)\,d\theta \\
&\quad = \int_{A\in\mathcal{R}_A} P\left(X^{1:n} \mid A\right) P(A)\,dA \cdot \int_{\Lambda\in\mathcal{R}_\Lambda} \Lambda\,P\left(Y^{1:n} \mid X^{1:n}, \Lambda\right) P(\Lambda)\,d\Lambda \\
&\quad = \frac{1}{\nu}\,\Phi\left(K^{(n)}\right)\widetilde{\Psi}\left(L^{(n)}\right)
\end{aligned}$$

for the parameters for the observation sequence, where

$$\widetilde{\Phi}\left(K^{(n)}\right) = \int_{A\in\mathcal{R}_A} A\,P\left(X^{1:n} \mid A\right) P(A)\,dA \quad \text{and}$$

$$\widetilde{\Psi}\left(L^{(n)}\right) = \int_{\Lambda\in\mathcal{R}_\Lambda} \Lambda\,P\left(Y^{1:n} \mid X^{1:n}, \Lambda\right) P(\Lambda)\,d\Lambda.$$

Here $\widetilde{\Phi}\left(K^{(n)}\right)$ is a $\nu \times \nu$ matrix, and $\widetilde{\Psi}\left(L^{(n)}\right)$ is a length ν vector, while $\Phi\left(K^{(n)}\right)$ and $\Psi\left(L^{(n)}\right)$ are scalars. Now, we can rewrite the BPM (2) as

$$\widehat{A} = E\left(A \mid Y^{1:n}\right) = \frac{\sum_{X^{1:n}\in\Omega_n} \widetilde{\Phi}\left(K^{(n)}\right)\Psi\left(L^{(n)}\right)}{\sum_{X^{1:n}\in\Omega_n} \Phi\left(K^{(n)}\right)\Psi\left(L^{(n)}\right)}$$

and

$$\widehat{\Lambda} = E\left(\Lambda \mid Y^{1:n}\right) = \frac{\sum_{X^{1:n}\in\Omega_n} \Phi\left(K^{(n)}\right)\widetilde{\Psi}\left(L^{(n)}\right)}{\sum_{X^{1:n}\in\Omega_n} \Phi\left(K^{(n)}\right)\Psi\left(L^{(n)}\right)}. \tag{9}$$

But, as for the estimate $\widehat{A}$, we see that the only difference between $\widetilde{\Phi}(K^{(n)})$ and $\Phi(K^{(n)})$ is the extra A inside the integration; i.e., considering $\widehat{A}$ elementwise, to obtain $\hat{a}_{ij} = E(a_{ij})$, we simply need to increment $k_{ij}^{(n)} \in K^{(n)}$ by one, due to the extra a_{ij} in the integration. So, let E_{ij} denote the $\nu \times \nu$ matrix with a 1 in entry (i, j) and 0 in every other entry, where the indices of the rows and columns start from 0 as with A. Thus, since A is a probability matrix, we get $\widehat{A} = \{\hat{a}_{ij}\}_{i,j \in \zeta_X}$ by letting

$$\hat{a}_{ij} = \frac{\sum_{X^{1:n} \in \Omega_n} \Phi\left(K^{(n)} + E_{ij}\right) \Psi\left(L^{(n)}\right)}{\sum_{X^{1:n} \in \Omega_n} \Phi\left(K^{(n)}\right) \Psi\left(L^{(n)}\right)} \text{ for } j \neq i-1 \quad \text{and}$$

$$\hat{a}_{i,i-1} = 1 - \sum_{j \in \zeta_X, j \neq i-1} \hat{a}_{ij}. \tag{10}$$

3.3 Partitioning Ω_n to get $\widehat{\Omega}_n$

Now to significantly reduce the computational complexity, the idea is to switch the summation from over all the possible Markov chain realizations Ω_n to over all the possible complete data sufficient statistics, the counts $(K^{(n)}, L^{(n)})$. For this, we partition Ω_n into equivalent classes with respect to $(K^{(n)}, L^{(n)})$. First, define an integer $\rho = \rho\left(Y^{1:n}\right)$ as

$$\rho = \max\left\{u \mid u \in Y^{1:n}\right\} + 1. \tag{11}$$

Here, we are not setting the upper bound for the range ζ_Y but assigning the label ρ to an integer that exists for finite observed sequences. The algorithm goes sequentially from time $t = 1$ to $t = n$; and rather than going through distinct Markov sequences, it goes through distinct values of $(K^{(n)}, L^{(n)})$. So, we define ω_t as

$$\omega_t = \left(K^{(t)},\, L^{(t)},\, x_1,\, x_t\right),$$

where $K^{(t)} = \{k_{ij}^{(t)}\}_{i,j \in \zeta_X}$ and $L^{(t)} = \{l_{iu}^{(t)}\}_{i \in \zeta_X, u \in \zeta_Y}$ are as defined in (3) and (4), and t is any positive integer, $1 \leq t \leq n$. Here ω_t can be viewed as a long vector of length $\nu \times \nu + \nu \times \rho + 2$, though for the notational convenience the elements are written in the form of $\nu \times \nu$ and $\nu \times \rho$ matrices followed by 2 integers.

The key is that, given a particular observation sequence, except for a few special cases, more than one distinct state sequence $X^{1:t} \in \Omega_t$ corresponds to the same value of ω_t, hence to the same value of $\Phi\left(K^{(t)}\right)$, $\Psi\left(L^{(t)}\right)$. So, we first construct the equivalence classes. It is obvious that each Markov sequences in Ω_t can be labeled and uniquely identified so that we can write $\Omega_t = \{X_1^{1:t}, X_2^{1:t}, \ldots, X_{\nu^t}^{1:t}\}$ with no ambiguity (e.g., by writing the states from left to right then reading it as a number written in base ν). Let $\mathcal{M}$ be a function such that

$$\mathcal{M}(X^{1:t}, Y^{1:t}) = \omega_t;$$

i.e., $\mathcal{M}$ returns the counts made on particular combination of $X^{1:t}$ and $Y^{1:t}$ plus the first and the last states. Now we define equivalence classes $[\,\cdot\,]$ for a fixed $Y^{1:t}$ as

$$\left[X_r^{1:t}\right] = \{X_s^{1:t} \in \Omega_t \mid \mathcal{M}(X_s^{1:t}, Y^{1:t}) = \mathcal{M}(X_r^{1:t}, Y^{1:t})\}$$

for any $s, r \in \{1, 2, \ldots, \nu^t\}$. In other words, $X_s^{1:t}$ and $X_r^{1:t}$ are in the same class if and only if the counts $(K^{(t)}, L^{(t)})$ and the first and last states are the same. We want to find the cardinality of the class $[X_r^{1:t}]$, which is more than one most of the time. Since the classes are defined via the value of ω_t, there is a one-to-one correspondence between ω_t-values in $\widehat{\Omega}_t$ and the equivalence classes $[X_r^{1:t}]$. The notation we now need is for a function that returns the cardinality of the equivalence class that corresponds to the given ω_n. So, again for a fixed $Y^{1:t}$, define a function $\mathcal{H}_t$ as

$$\mathcal{H}_t(\omega_t) = \left\{\text{the cardinality of } \left[X_r^{1:t}\right] \text{ for which } \mathcal{M}\left(X_r^{1:t}, Y^{1:t}\right) = \omega_t\right\}.$$

Furthermore, define $\widehat{\Omega}_t$ as the set of all the possible ω_t-values; i.e.,

$$\widehat{\Omega}_t = \{\omega_t \mid \mathcal{M}\left(X^{1:t}, Y^{1:t}\right) = \omega_t \text{ for some } X^{1:t} \in \Omega_t\}.$$

Using $\mathcal{H}_n(\omega_n)$ and $\widehat{\Omega}_n$, we rewrite the formula for the BPM so that the summations are not over Ω_n but over $\widehat{\Omega}_n$. From (10) we have a matrix $\widehat{A} = \{\hat{a}_{ij}\}_{i,j\in\zeta_X}$ such that

$$\hat{a}_{ij} = \frac{\sum_{\omega_n\in\widehat{\Omega}_n} \mathcal{H}_n(\omega_n)\,\Phi\left(K^{(n)} + E_{ij}\right)\Psi\left(L^{(n)}\right)}{\sum_{\omega_n\in\widehat{\Omega}_n} \mathcal{H}_n(\omega_n)\,\Phi\left(K^{(n)}\right)\Psi\left(L^{(n)}\right)} \text{ for } j \neq i-1 \quad \text{and}$$

$$\hat{a}_{i,i-1} = 1 - \sum_{j\,\in\,\zeta_X,\, j\neq i-1} \hat{a}_{ij},$$

while from (9) we have a vector $\widehat{\Lambda} = \left(\widehat{\lambda}_0, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_{\nu-1}\right)$ such that

$$\widehat{\Lambda} = \frac{\sum_{\omega_n\in\widehat{\Omega}_n} \mathcal{H}_n(\omega_n)\,\Phi\left(K^{(n)}\right)\widetilde{\Psi}\left(L^{(n)}\right)}{\sum_{\omega_n\in\widehat{\Omega}_n} \mathcal{H}_n(\omega_n)\,\Phi\left(K^{(n)}\right)\Psi\left(L^{(n)}\right)}, \tag{12}$$

It is obvious, considering the range of ω_n, the size of $\widehat{\Omega}_n$ is polynomial in n and significantly smaller than the size of Ω_n, ν^n, except for very small values of n. The next section will show that the computational complexity remains polynomial in n through an algorithm finding $\widehat{\Omega}_n$ and $\mathcal{H}_n$-values.

3.4 Algorithm to Find $\widehat{\Omega}_n$

Algorithm 1 finds $\widehat{\Omega}_n$ and $\mathcal{H}_n(\omega_n)$. It is an extension of the previous version (See Murakami and Taylor 2006) after sacrificing rather an insignificant amount of computational complexity for the simplicity (i.e., we could use a vector of smaller dimension than ω_n to find ($K^{(n)}$, $L^{(n)}$), as shown for the two-state case described in Section 3.5).

Starting from $t = 1$, the sequential algorithm below finds all the values of $\mathcal{H}_t(\omega_t)$ for all $\omega_t \in \widehat{\Omega}_t$, $t = 1, 2, \ldots, n$, given an observation sequence $Y^{1:n}$. In each step, it processes an element ω_{t-1} picked from the set $\widehat{\Omega}_{t-1}$, considering all the possible value of x_t and modifying ω_{t-1} to get ω_t for each of the x_t-value so that ω_t is logically true when the current x_t-value is added to $X^{1:t-1}$ (Line 6), and put it in the set $\widehat{\Omega}_t$ (here more than one element in $\widehat{\Omega}_{t-1}$ often end up being modified into the same element in $\widehat{\Omega}_t$), while updating $\mathcal{H}_t\left(\omega_t\right)$ each time by assigning $\mathcal{H}_{t-1}\left(\omega_{t-1}\right)$ if ω_t is not in $\widehat{\Omega}_t$ yet or by adding $\mathcal{H}_{t-1}\left(\omega_{t-1}\right)$ to it if otherwise (Lines 7 through 10). Note, with a naive approach, an algorithm to find $\widehat{\Omega}_n$ and $\mathcal{H}_n(\omega_n)$ itself could still have exponential computational complexity in n, which we avoid here.

One minor issue can be taken care of to reduce the complexity a little further. Let $\widehat{\Omega}_t|_{x_1=i}$ be a partition of $\widehat{\Omega}_t$, $i \in \zeta_X$, such that

$$\widehat{\Omega}_t|_{x_1=i} = \left\{\omega_t \in \widehat{\Omega}_t \mid \omega_t = \left(K^{(t)},\, L^{(t)},\, i,\, x_t\right)\right\}.$$

Instead of working on all the elements in $\widehat{\Omega}_t$ as described above, we could just work on the elements in $\widehat{\Omega}_t|_{x_1=0}$ instead, which can be achieved by just putting ω_1 that corresponds to the case $X^{1:1} = x_1 = 0$ and no other elements into $\widehat{\Omega}_1|_{x_1=0}$ (Line 1). It is because once we find $\widehat{\Omega}_n|_{x_1=0}$ and all the corresponding $\mathcal{H}_n$-values, we can find the rest ($\widehat{\Omega}_n|_{x_1=i}$ for all $i > 0$, $i \in \zeta_X$, and the corresponding $\mathcal{H}_n$-values) by interchanging every 0 and i that appear in the subscripts representing a state, in a way described below.

Let $\psi(\omega_n, i)$ be the ω_n-value that is obtained from $\omega_n \in \widehat{\Omega}_1|_{x_1=0}$ by interchanging the states 0 and i, where $i \in \zeta_X \setminus \{0\}$. This $\psi(\omega_n, i)$ can be described using three functions $\psi_{_K}$, $\psi_{_L}$, and ψ_x defined as follows. With the row and column indices starting at 0, let $\psi_{_K}(K, i)$ be the matrix that is obtained by interchanging the i-th row with the 0-th row and the i-th column with the 0-th column of a $\nu \times \nu$ matrix K given, let $\psi_{_L}(L, i)$ be the matrix that is obtained by interchanging the i-th row with the 0-th row of a $\nu \times \rho$ matrix L given, and let $\psi_x(x, i)$ be 0 if $x = i$, i if $x = 0$, and x if otherwise for $x \in \zeta_X$. Finally, define $\psi(\omega_n, i)$ on $\omega_n = (K^{(n)},\, L^{(n)},\, 0,\, x_n) \in \widehat{\Omega}_n|_{x_1=0}$ and $i \in \zeta_X \setminus \{0\}$ as

$$\begin{aligned}\psi(\omega_n, i) &= \psi\left((K^{(n)},\, L^{(n)},\, 0,\, x_n),\, i\right)\\ &= \left(\psi_K\left(K^{(n)}, i\right),\, \psi_L\left(L^{(n)}, i\right),\, i,\, \psi_x(x_n, i)\right);\end{aligned}$$

Then, for any $\omega_n \in \widehat{\Omega}_n|_{x_1=0}$ and any $i \in \zeta \setminus \{0\}$, we have

$$\psi(\omega_n, i) \in \widehat{\Omega}_n|_{x_1=i} \quad \text{and} \quad \mathcal{H}_n(\psi(\omega_n, i)) = \mathcal{H}_n(\omega_n).$$

Algorithm 1 *Find $\widehat{\Omega}_n|_{x_1=0}$ and its $\mathcal{H}_n$-values.*

```
Require: $E^{(m_1,m_2)}_{ij}$ is the $m_1 \times m_2$ matrix with a 1 in entry $(i, j)$ and 0 in
    every other entry, counting the rows and columns from 0.
let $\mathcal{H}_1(\omega_1) = 1$ and $\widehat{\Omega}_1|_{x_1=0} = \{\omega_1\}$, where $\omega_1 = (K^{(1)}, L^{(1)}, x_1, x_1) =$
    $(K_0, E^{(\nu,\rho)}_{0,y_t}, 0, 0)$, and $K_0$ is a $\nu \times \nu$ zero matrix
for $t = 2$ to $n$ do
    let $\widehat{\Omega}_t|_{x_1=0}$ be an empty set
    for all $\omega_{t-1}$ such that $\omega_{t-1} = (K^{(t-1)}, L^{(t-1)}, 0, x_{t-1}) \in \widehat{\Omega}_{t-1}|_{x_1=0}$ do
        for all $i$ such that $i \in \zeta_X$ do
            let $\omega_t = (K^{(t-1)} + E^{(\nu,\nu)}_{x_{t-1},i}, L^{(t-1)} + E^{(\nu,\rho)}_{i,y_t}, 0, i)$
            if $\omega_t$ is not in $\widehat{\Omega}_t|_{x_1=0}$ yet then
                let $\mathcal{H}_t(\omega_t) = \mathcal{H}_{t-1}(\omega_{t-1})$, and put $\omega_t$ in $\widehat{\Omega}_t|_{x_1=0}$
            else
                let $\mathcal{H}_t(\omega_t) = \mathcal{H}_t(\omega_t) + \mathcal{H}_{t-1}(\omega_{t-1})$
```

We see the upper bound for the complexity is at most $cn^{\nu(\nu+\rho)+3}$ for some constant c; i.e., the computational complexity is polynomial in n, while exponential in ν and ρ. As for the actual complexity, which depends on $Y^{1:n}$, some examples are given in Section 5.2, indicating an extremely small c.

3.5 2×2 chain

In this section we describe the algorithm for the case the state space size $\nu = 2$, which can be, of course, obtained by simply substituting ν by 2 in Algorithm 1. However, because of identities that exist among the elements in ω_t, for any ν, we could find the sufficient statistics $(K^{(n)},\, L^{(n)})$ using a 'smaller sized ω_t' and modifying Algorithm 1 accordingly (Murakami & Taylor, 2006). Although, with a very large ν, both the identities and the modified algorithm become rather complicated and hence not recommended to be used, with $\nu = 2$ it is relatively easy to implement as shown below.

In place of $\omega_t = (K^{(t)}, L^{(t)}, x_1, x_t)$ we use $\widetilde{\omega}_t$, which consists of minimal data necessary to obtain ω_t but is in smaller size than ω_t, where $\widetilde{\omega}_t$ is defined as

$$\widetilde{\omega}_t = \left(k_1^{(t)},\, k_{11}^{(t)},\, L_{\rho-1}^{(t)},\, x_1,\, x_t\right),$$

while we let, for $i \in \{0,1\}$,

$$k_i^{(t)} = \sum_{s=1}^{t} \mathbf{I}\left\{x_s = i, x_s \in X^{1:t}\right\}, \tag{13}$$

$$l_u^{(t)} = \sum_{s=1}^{t} \mathbf{I}\left\{y_s = u, y_s \in T^{1:t}\right\}, \quad \text{and} \tag{14}$$

$$L_{\rho-1}^{(t)} = \left(l_{11}^{(t)},\, l_{12}^{(t)},\, \ldots,\, l_{1,\rho-1}^{(t)}\right). \tag{15}$$

The vector $L_{\rho-1}^{(t)}$ is just $L^{(t)}$ without the first (or 0-th) row and column, and $\widetilde{\omega}_t$ is a vector of length $2 + (\rho - 1) + 2 = \rho + 3$. Now $\widetilde{\omega}_t$ has the minimal information necessary to find $(K^{(n)}, L^{(n)})$ and also to carry out the algorithm. (Note $l_u^{(n)}$-values are fixed once a particular $Y^{1:n}$ is given.) The identities below show how $(K^{(n)}, L^{(n)})$ can be obtained from $\widetilde{\omega}_n$. Letting $k_{ij} = k_{ij}^{(n)}$, $l_{iu} = l_{iu}^{(n)}$, $k_1 = k_1^{(n)}$, and $l_u = l_u^{(n)}$, and for $i, j \in \{0,1\}$ and $u \in \zeta_Y \setminus \{0\}$,

$$\begin{aligned}
&k_{10} = k_1 - k_{11} - x_n, \quad k_{01} = k_1 - k_{11} - x_1,\\
&k_{00} = n - 1 - k_{10} - k_{01} - k_{11} (= n - 1 - 2k_1 + k_{11} + x_1 + x_n),\\
&l_{10} = k_1 - \sum_{u=1}^{\rho-1} l_{1u}, \quad l_{0u} = l_u - l_{1u}, \quad \text{and} \quad l_{00} = n - \sum_{u=1}^{\rho-1} l_u - l_{10}.
\end{aligned}$$

Abusing the notation a little for easier comparison, we keep $\widetilde{\Omega}_n$ as the range of $\widetilde{\omega}_t$ (not of ω_t), given a particular $Y^{1:n}$. As before, to obtain $\widehat{\Omega}_n|_{x_1=1}$ out of $\widehat{\Omega}_n|_{x_1=0}$, we simply interchange the states 0 and 1 as described below.

Using the identities above, define $\psi_{k_{11}}$, $\psi_{L_{\rho-1}}$, and ψ_x as

$$\begin{aligned}
\psi_{k_{11}}(k_1, k_{11}, x_n) &= n - 1 - 2k_1 + k_{11} + x_n,\\
\psi_{L_{\rho-1}}(L_{\rho-1}) &= (l_1, l_2, \ldots, l_{\rho-1}) - L_{\rho-1}, \quad \text{and}\\
\psi_x(x) &= \begin{cases} 1 & \text{if } x = 0\\ 0 & \text{if } x = 1. \end{cases}
\end{aligned}$$

Then, define ψ on $\widehat{\Omega}_n|_{x_1=0}$ as

$$\psi(k_1,\, k_{11},\, L_{\rho-1},\, 0,\, x_n) = \left(n - k_1,\, \psi_{k_{11}}(k_1, k_{11}, x_n),\, \psi_{L_{\rho-1}}(L_{\rho-1}),\, 1,\, \psi_x(x_n)\right).$$

Now, for any $\widetilde{\omega}_n = \left(k_1^{(n)},\, k_{11}^{(n)},\, L_{\rho-1}^{(n)},\, 0,\, x_n\right) \in \widehat{\Omega}_n|_{x_1=0}$, we have

$$\psi(\widetilde{\omega}_n) \in \widehat{\Omega}_n|_{x_1=1} \quad \text{and} \quad \mathcal{H}_n\left(\psi(\widetilde{\omega}_n)\right) = \mathcal{H}_n(\widetilde{\omega}_n).$$

Algorithm 2 shows how it is done if $\widetilde{\omega}_t$ is used instead of ω_t. The vector V_u defined in the algorithm corresponds to $E_{iu}^{(2,\rho)}$ in Algorithm 1. Note the value $i \in \{0,1\}$ for the state x_t is also used as a delta function in Line 6, which corresponds to $E_{x_{t-1},i}^{(\nu,\nu)}$ and $E_{i,y_t}^{(\nu,\rho)}$, $\nu = 2$, in Algorithm 1.

Algorithm 2 *Find $\widehat{\Omega}_n|_{x_1=0}$ and its $\mathcal{H}_n$-values for $\nu = 2$ (previous version).*

```
Require: $V_u$ is the length $\rho - 1$ vector with $u$-th entry 1 and 0 in every other
    entry, counting the entries from 1.
 1: let $\mathcal{H}_1(\widetilde{\omega}_1) = 1$ and $\widehat{\Omega}_1|_{x_1=0} = \{\widetilde{\omega}_1\}$, where $\widetilde{\omega}_1 = \left(k_1^{(1)},\, k_{11}^{(1)},\, L_{\rho-1}^{(1)},\, 0,\, x_1\right)$ is
    a length-$\rho + 3$ zero vector
 2: for $t = 2$ to $n$ do
 3:     let $\widehat{\Omega}_t|_{x_1=0}$ be an empty set
 4:     for all $\widetilde{\omega}_{t-1}$ such that $\widetilde{\omega}_{t-1} = \left(k_1^{(t-1)},\, k_{11}^{(t-1)},\, L_{\rho-1}^{(t-1)},\, 0,\, x_{t-1}\right) \in$
        $\widehat{\Omega}_{t-1}|_{x_1=0}$ do
 5:         for all $i$ such that $i \in \{0,1\}$ do
 6:             let $\widetilde{\omega}_t = \left(k_1^{(t-1)} + i,\, k_{11}^{(t-1)} + x_{t-1} \cdot i,\, L_{\rho-1}^{(t-1)} + V_{y_t} \cdot i,\, 0,\, i\right)$
 7:             if $\widetilde{\omega}_t$ is not in $\widehat{\Omega}_t|_{x_1=0}$ yet then
 8:                 let $\mathcal{H}_t(\widetilde{\omega}_t) = \mathcal{H}_{t-1}(\widetilde{\omega}_{t-1})$, and put $\widetilde{\omega}_t$ in $\widehat{\Omega}_t|_{x_1=0}$
 9:             else
10:                 let $\mathcal{H}_t(\widetilde{\omega}_t) = \mathcal{H}_t(\widetilde{\omega}_t) + \mathcal{H}_{t-1}(\widetilde{\omega}_{t-1})$
```

As for the integration over A, after replacing the elements of D in $\Phi_{_D}(D)$ that appears in Equations (A.1) and (A.2a) through (A.2h) in Appendix A with the corresponding k_{ij}, we find $\Phi\left(K^{(n)}\right)$ as

$$\Phi\left(K^{(n)}\right) = \sum_{i=0}^{k_{10}} \frac{(-1)^i k_{10}!}{i!(k_{10}-i)!(k_{11}+i+1)} \cdot \frac{(k_{00}+k_{11}+i+1)!\,k_{01}!}{(n-k_{10}+i+1)!}.$$

In the above, the identity $\sum_{i=0}^{1}\sum_{j=0}^{1} k_{ij} = n - 1$ is used to simplify the second denominator. Define

$$\phi\left(i,\, K^{(n)}\right) = \frac{(-1)^i k_{10}!}{i!(k_{10}-i)!(k_{11}+i+1)} \cdot \frac{(k_{00}+k_{11}+i+1)!\,k_{01}!}{(n-k_{10}+i+1)!}$$

so that $\Phi\left(K^{(n)}\right) = \sum_{i=0}^{k_{10}} \phi\left(i,\, K^{(n)}\right)$. By incrementing the corresponding exponent k_{ij} for $\hat{a}_{ij}$ by 1, we get the estimates as shown below.

$$\hat{a}_{00} = \frac{1}{2P(Y^{1:n})} \sum_{\widetilde{\omega}_n \in \widehat{\Omega}_n} \mathcal{H}_n(\widetilde{\omega}_n) \left[\sum_{i=0}^{k_{10}} \frac{k_{00}+k_{11}+i+2}{n-k_{10}+i+2} \phi\Big(i,\, K^{(n)}\Big)\right] \Psi\Big(L^{(n)}\Big) \quad \text{and}$$

$$\hat{a}_{11} = \frac{1}{2P(Y^{1:n})} \sum_{\widetilde{\omega}_n \in \widehat{\Omega}_n} \mathcal{H}_n(\widetilde{\omega}_n) \left[\sum_{i=0}^{k_{10}} \frac{k_{11}+i+1}{k_{11}+i+2}\, \frac{k_{00}+k_{11}+i+2}{n-k_{10}+i+2} \phi\Big(i,\, K^{(n)}\Big)\right] \cdot \Psi\Big(L^{(n)}\Big),$$

where $P(Y^{1:n}) = \dfrac{1}{2} \sum_{\widetilde{\omega}_n \in \widehat{\Omega}_n} \mathcal{H}_n(\widetilde{\omega}_n)\, \Phi(K^{(n)})\, \Psi\left(L^{(n)}\right)$. Note, E_{ij} used in (10) is not used in these equations because we rather want to utilize $\phi\left(i,\ K^{(n)}\right)$-values that are already found for $P(Y^{1:n})$.

4 Poisson Hidden Markov Models

Consider a discrete-time HMM such that the observation sequence is governed by Poisson distributions with parameter values depending on the current state. Let $f_{i,\lambda_i}(u)$ be the density functions such that, for $i \in \zeta_X$ and $u \in \zeta_Y = \{0, 1, \ldots, \infty\}$,

$$f_{i,\lambda_i}(u) = \frac{\lambda_i^u e^{-\lambda_i}}{u!} = P\left(y_t = u \mid x_t = i\right) \quad \text{for any integer } t > 0,$$

and let $\Lambda = \{\lambda_i\}_{i \in \zeta_X}$; i.e., the Poisson parameter depends on the current state.

4.1 Bayesian Posterior Mean Estimates for PHMM

As for the integration over A, refer to Appendix A. As for the integration over Λ, since the symmetry in the distribution is already taken care of in the integration over A, it is simply over the entire domain of Λ. So, $\Psi\left(L^{(n)}\right)$ in (6) can be rewritten as a product of integrals. As for the prior $P(\Lambda)$, in this paper, we choose independent gamma distributions for each parameter; i.e.,

$$\begin{aligned} P(\Lambda) &= \prod_{i \in \zeta_X} P(\lambda_i) \quad \text{where} \\ P(\lambda_i) &= f(\lambda_i; \alpha_i, \beta_i) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_i^{\alpha_i - 1} e^{-\beta_i \lambda_i}. \end{aligned}$$

Other types of prior distributions could be chosen here with minor modifications on the formula shown below, accordingly. With $k_i = k_i^{(n)}$, $l_u = l_u^{(n)}$, and $l_{iu} = l_{iu}^{(n)}$ as defined in (13), (14), and (4), respectively, we have

$$\begin{aligned} \Psi\left(L^{(n)}\right) &= \prod_{i \,\in\, \zeta_X} \left[\int_0^\infty \prod_{u \,\in\, \zeta_Y} \left(\frac{\lambda_i^u e^{-\lambda_i}}{u!} \right)^{l_{iu}} \left(\frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_i^{\alpha_i - 1} e^{-\beta_i \lambda_i} \right) d\lambda_i \right] \\ &= \frac{N_0}{N_1} \prod_{i \in\, \zeta_X} \int_0^\infty \lambda_i^{N_{2i} - 1} \exp\left(-N_{3i} \lambda_i\right) d\lambda_i \\ &= \frac{N_0}{N_1} \prod_{i=0}^{\nu - 1} \frac{\Gamma(N_{2i})}{N_{3i}^{N_{2i}}}, \end{aligned}$$

where

$$N_0 = \prod_{i \in \zeta_X} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}, \quad N_1 = \prod_{u \in \zeta_Y} (u!)^{l_u}, \quad N_{2i} = \alpha_i + \sum_{u \in \zeta_Y} u\, l_{iu}, \quad \text{and}$$
$$N_{3i} = \beta_i + k_i. \tag{16}$$

As for N_1, N_{2i} and N_{3i}, if u does not appear in $Y^{1:n}$, we have $l_{iu} = 0$; so, disregarding the factors and/or terms that involve such zero-valued l_{iu} will have no effect on the final values of these three. So, in above the multiplications and/or summations that are over ζ_Y in N_1, N_{2i}, and N_{3i} are actually over the finite subset $\{0, 1, \ldots, \rho - 1\} \subset \zeta_Y$, where $\rho = \max\{u \mid u \in Y^{1:n}\} + 1$ is as defined in (11).

Regarding the estimate of λ_i, multiplying an extra λ_i inside the integral of $\Psi\left(L^{(n)}\right)$ will result only in increasing the value of N_{2i} by 1. Hence, $\widetilde{\Psi}\left(L^{(n)}\right)$ in (12) is a length-ν vector such that the i-th element, say $\widetilde{\Psi}_i\left(L^{(n)}\right)$, is obtained by

$$\widetilde{\Psi}_i\left(L^{(n)}\right) = \frac{N_{2i}}{N_{3i}} \cdot \Psi\left(L^{(n)}\right), \tag{17}$$

and this $\widetilde{\Psi}_i\left(L^{(n)}\right)$ is used to obtain the estimate of λ_i as

$$\widehat{\lambda}_i = \frac{\sum_{\omega_n \in \widehat{\Omega}_n} \mathcal{H}_n(\omega_n)\, \Phi\left(K^{(n)}\right) \widetilde{\Psi}_i\left(L^{(n)}\right)}{\sum_{\omega_n \in \widehat{\Omega}_n} \mathcal{H}_n(\omega_n)\, \Phi\left(K^{(n)}\right) \Psi\left(L^{(n)}\right)},$$

where $\Phi\left(K^{(n)}\right)$ is as shown in Appendix A.

As for the values of $\mathcal{H}_n(\omega_n)$ and the estimates for the transition matrix A, see Sections 3.3 and 3.4.

4.2 2 × 2 Chain for PHMM

For the case $\nu = 2$, we have $\Lambda = \{\lambda_1, \lambda_2\}$. As for the BPM for Λ, from (16), we get

$$\Psi\left(L^{(n)}\right) = \frac{N_0}{N_1} \frac{\Gamma(N_{20})}{N_{30}^{N_{20}}} \frac{\Gamma(N_{21})}{N_{31}^{N_{21}}}.$$

where N_0, N_1, N_{2i}, and N_{3i}, $i = 0, 1$, are as defined also in (16). Also, $\widetilde{\Psi}\left(L^{(n)}\right)$ is now $\left(\widetilde{\Psi}_0(L^{(n)}), \widetilde{\Psi}_1(L^{(n)})\right)$ and can be obtained by (17) above. As for the values of $\mathcal{H}_n(\omega_n)$ and the estimates for the transition matrix A, see Section 3.5.

5 Results

One of the well-known problems of the MLE occurs when the estimate $\hat{\theta}$ that maximizes the likelihood, given a particular observation sequence, is not necessarily close to the true parameter set θ especially when the data size is small. In order to consider a similar problem in the MLE for PHMMs, root mean square errors for the MLE and BPM are obtained in the way described below. In addition, to approximate the computational complexity of the BPM, the distribution of the size of $\widehat{\Omega}_n$ is shown by simulations.

Since up to 1200 estimations are to be computed for each, to make the simulations feasible for the algorithm, only λ_0- and λ_1-values such that $0 < \lambda_0, \lambda_1 \leq 3$ are considered hereafter.

As for the BPM, for all $i, j \in \{0, 1\}$, the parameter values for the prior $P(\Lambda)$ are set as $\alpha_i = 1.5$ and $\beta_i = 0.5$ (see Section 4.1), while the parameter values for the prior $P(A)$ are set as $\gamma_{ij} = 1$ (see Appendix A).

As for the MLE, an EM algorithm (Dempster et al., 1977) obtained by a minor modification to the standard Baum-Welch algorithm (Baum et al., 1970; Baum, 1972; Rabiner, 1989) is used in the following way: first obtain 15 estimates (or fixed points) using 15 randomly picked initial estimates, then choose the estimate that gives the highest likelihood as the final estimate. In case more than one give similar likelihood, choose the one with the largest basin of convergence as the final estimate.

It is a well-known limitation of the Baum-Welch algorithm that the algorithm is not guaranteed to find the estimate that gives the absolute maximum and has a dependency on the initial estimate, which also applies to this modified version. However, empirically we see that the smaller the data size is, the smoother the likelihood surface hence the larger basins of conversion. With n and the number of initial estimates used for the simulation, the EM estimates obtained are highly likely the MLE most of the time, and an increase in the number of initial estimates from 15 did not noticeably changed the outcome. (As a reference for the EM method used here, with the data size $n = 1000$, using randomly generated 1000 PHMMs, the root mean square errors in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ had the mean 0.40, the median 0.20, and the variance 0.47, while the best estimations were obtained when $|\lambda_0 - \lambda_1| \in (2, 3]$ and $\det(A) \in [-1, -0.8]$, with the mean 0.07, the median 0.07, and the variance 0.05.)

5.1 Root Mean Square Errors w.r.t. $|\lambda_0 - \lambda_1|$ and the Determinant of A

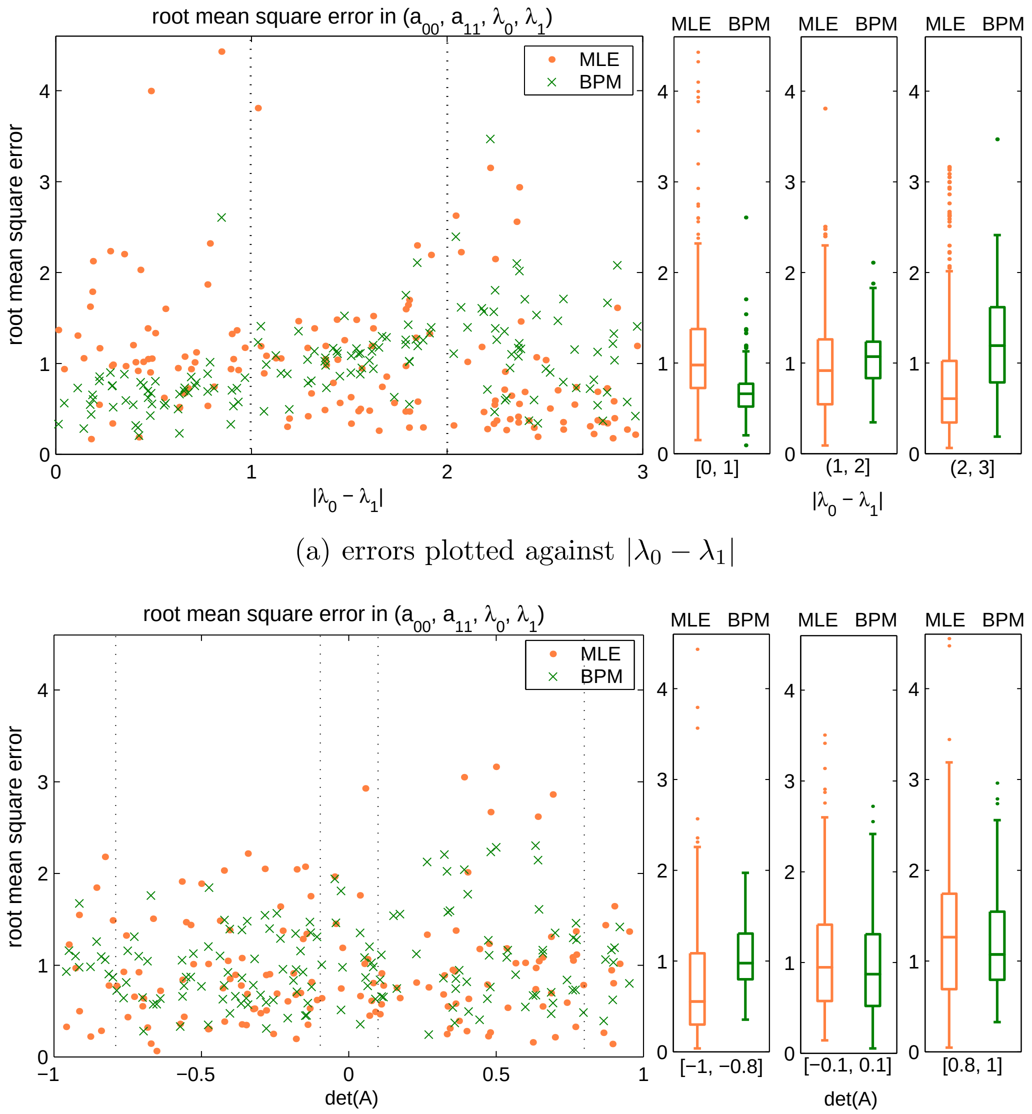
The range of $|\lambda_0 - \lambda_1|$, which is set as $[0, 3]$ for simulations, is partitioned into three subintervals: $[0, 1]$, $(1, 2]$, and $(2, 3]$. Then, for each subintervals, 400 θ-values are picked randomly with respect to both $|\lambda_0 - \lambda_1|$ and the determinant of A. Using each of these θ-values, an observation sequence of length $n = 30$ is generated and two estimates, the MLE and BPM, are found.

A similar experiment is implemented again but with three subintervals of the determinant of A: $[-1, -0.8]$, $[-0.1, 0.1]$, and $[0.8, 1]$ (which obviously do not sum up to the entire range, $[-1, 1]$, but are chosen simply to emphasize the difference). The interval $[-1, -0.8]$ is for the case both a_{00} and a_{11} are very small, resulting in quickly alternating state sequences. The interval $[-0.1, 0.1]$ is for the opposite case, both a_{00} and a_{11} are close to 1; the state tends to stay the same a long time before switching to the other. Finally, the interval $[0.8, 1]$ is for the case $a_{0j} \approx a_{1j}$, $j \in \{0, 1\}$; i.e., the current state does not have much effect on the probability distribution for the next states. For example, if $a_{00} \approx a_{10} \approx 1$ then the state sequence stays in 0 most of the time, with occasional rare appearances of a single 1. On the other hand, if $a_{00} \approx a_{10} \approx 0.5$, the sequence alternates almost randomly. Again, 400 θ-values are randomly picked for each interval so that it is uniformly distributed with respect to $|\lambda_0 - \lambda_1|$ and to each of the subintervals of $\det(A)$. (Incidentally, additional estimates are obtained for the scatter plot, which covers the entire range of $\det(A)$.)

Thus, in total $400 \times 3 = 1200$ PHMMs are generated for each simulation, and Figs. 1(a) and 1(b) show the results regarding the effect of $|\lambda_0 - \lambda_1|$ and $\det(A)$ on the root mean square errors in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$, respectively. The scatter plots are for 150 pairs of estimates in total, while the boxplots are for 400 pair of estimates per box. A few outliers are out of the frame for better views.

We can see some correlation in each of the plots; however, the correlations get a little more obvious if we plot the root mean square errors in (λ_0, λ_1) against $|\lambda_0 - \lambda_1|$ and in (a_{00}, a_{11}) against $\det(A)$ as shown in Figs. 2(a) and 2(b), respectively. Again, the scatter plots are for 150 pairs of estimates in total, while the boxplots are for 400 pair of estimates per box; and a few outliers are out of the frame.

Regarding $|\lambda_0 - \lambda_1|$, we see in Fig. 2(a) that, compared to the MLEs, the BPMs tend to be closer to the true values and much more stable when the difference $|\lambda_0 - \lambda_1|$ is relatively small; in other words, when the two states are close. These characteristics switch around, making the MLE the better choice, when $|\lambda_0 - \lambda_1|$ increases toward 3. This indicates that in the first case the global (and local) maximum of the likelihood surface is unstable and often



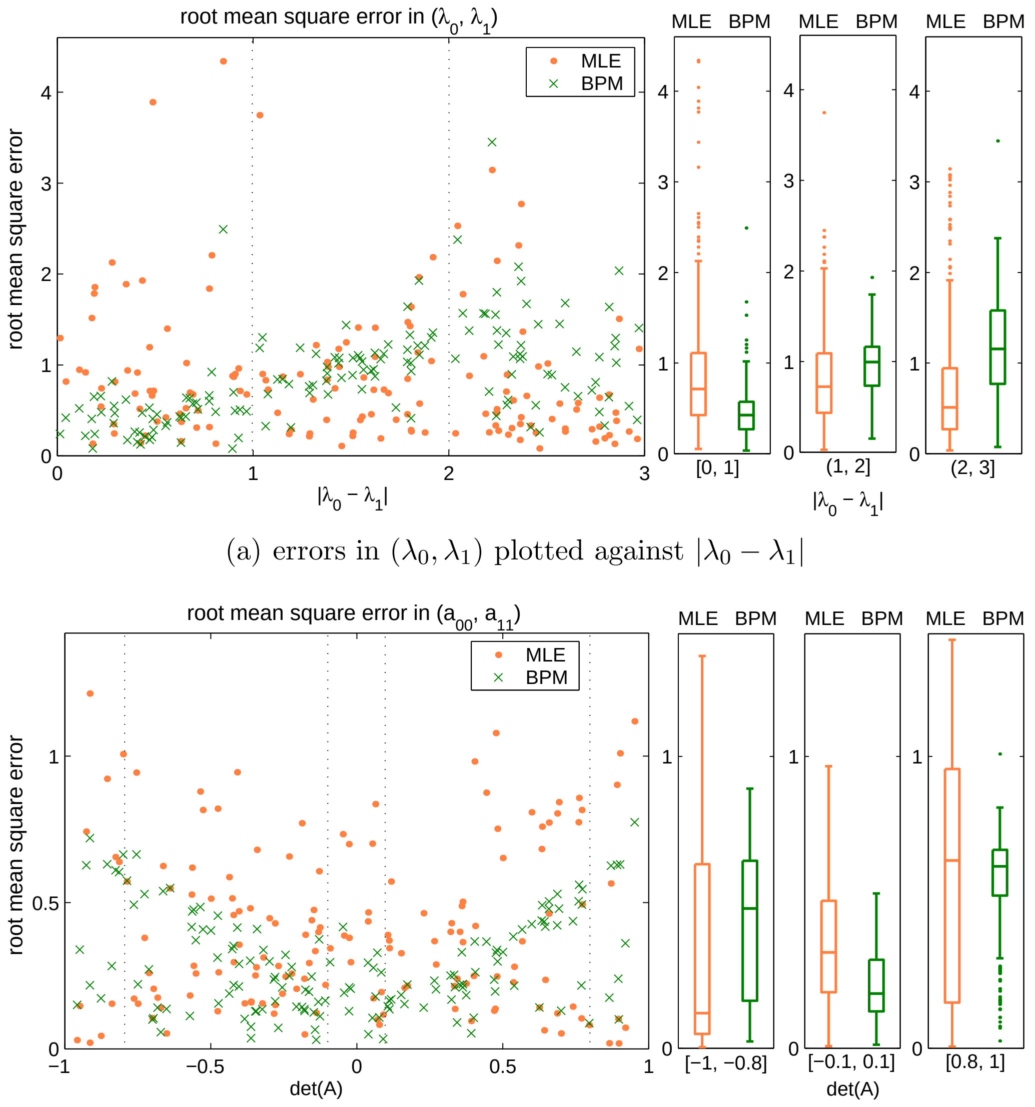
(a) errors plotted against $|\lambda_0 - \lambda_1|$

(b) errors plotted against the determinant of A

Fig. 1. The Root Mean Square Errors in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$, $n = 30$

not close to the true parameter set, which gives an advantage to the BPMs; and in the second case the global maximum tends to stay close to the true parameter set, while the distribution surface has a very long asymmetric tail, which gives an advantage to the MLEs and an instability to the BPMs.

Two examples of the likelihood surface are plotted against λ_0 and λ_1: one for when $|\lambda_0 - \lambda_1|$ is small (0.5) and the other for when it is larger (2.5) in Figs. 3 and 4, respectively, with $n = 40$. In order to visualize the likelihood distribution that is actually in five dimensions (including the likelihood itself and with the initial state distribution being fixed) in three dimensions, the values a_{00} and a_{11} are fixed to those of the MLE estimate in (a) and of the BPM



(a) errors in (λ_0, λ_1) plotted against $|\lambda_0 - \lambda_1|$

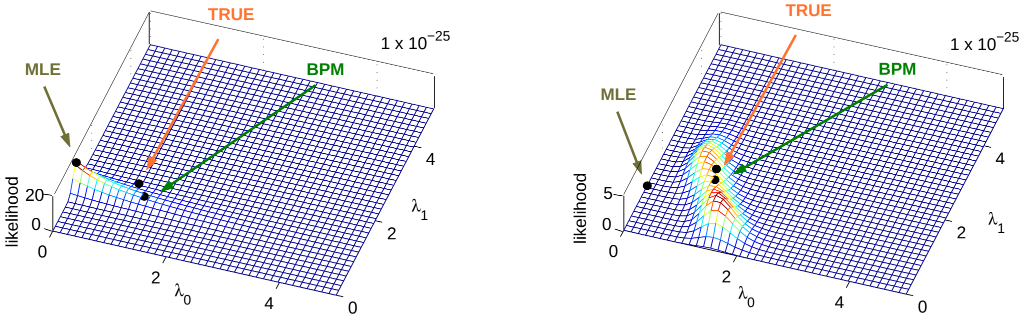
(b) errors in (a_{00}, a_{11}) plotted against $\det(A)$

Fig. 2. The Root Mean Square Errors in (λ_0, λ_1) and (a_{00}, a_{11}), $n = 30$

estimate in (b). Hence, they do not give the whole view (which is impossible) but provide some hints about it none the less.

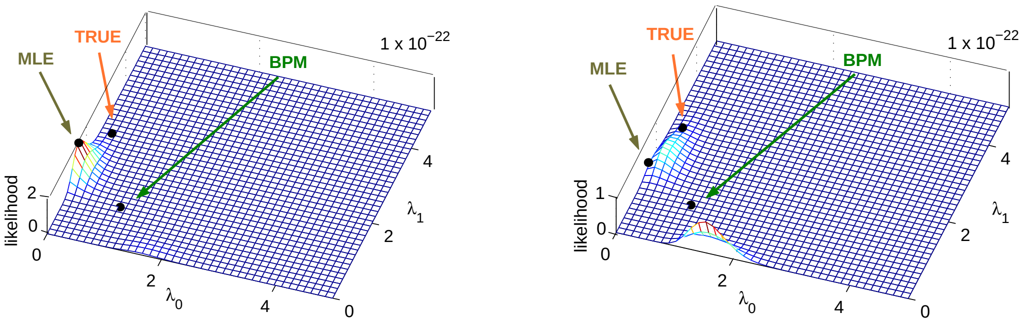
The parameter values $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ concerned are as follows: in Fig. 3 the true value is $(0.8, 0.6, 1.0, 1.5)$, the MLE is $(0.35, 1.00, 0.00, 1.20)$, and the BPM is $(0.63, 0.28, 1.19, 1.20)$; and in Fig. 4 the true value is $(0.8, 0.6, 0.2, 2.7)$, the MLE is $(0.56, 0.68, 0.00, 1.60)$, and the BPM is $(0.69, 0.46, 0.92, 1.02)$. For each estimate, from the pair of estimates in symmetry, the one with the smaller error in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ is plotted, and the determinant of A is $0.8 + 0.6 - 1 = 0.4$.

Regarding the effects of $\det(A)$, we see in Fig. 2(b) the difference is mainly in the stability. We see over-all higher stability of the BPMs compared to



(a) viewed at $(\hat{a}_{00}, \hat{a}_{11})$ of MLE (b) viewed at $(\hat{a}_{00}, \hat{a}_{11})$ of BPM

Fig. 3. Example of Likelihood Surface for $|\lambda_0 - \lambda_1| = 0.5$, $n = 40$



(a) viewed at $(\hat{a}_{00}, \hat{a}_{11})$ of MLE (b) viewed at $(\hat{a}_{00}, \hat{a}_{11})$ of BPM

Fig. 4. Example of Likelihood Surface for $|\lambda_0 - \lambda_1| = 2.5$, $n = 40$

the MLEs, while the performances of both estimates are better when the determinant of A is smaller (i.e., when the state stays the same relatively long in either state), which agrees with our intuition.

Considering the results above, we now focus on the combinations of the subintervals in which the error distributions of MLEs and BPMs are found to be considerably different, which are $[0, 1]$ and $(2, 3]$ for $|\lambda_0 - \lambda_1|$, and $[-1, -0.8]$ and $[0.8, 1]$ for $\det(A)$. Under the restriction $0 < \lambda_0, \lambda_1 \leq 3$ as before, 200 θ-values are randomly picked for each of all the possible combinations of these subintervals so that θ are uniformly distributed with respect to the intervals. Then for each of the θ-values picked, an observation sequence is generated with $n = 30$, and the BPM and MLE (using the EM estimator in the same way as before) are obtained. The root mean square error of the estimates in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ are shown in Fig. 5 and in Table 1.

As expected, when the change in Poisson parameter is small (i.e., when $|\lambda_0 - \lambda_1| \in [0, 1]$), the change in the Markov chain characteristics (as far as we see by $\det(A)$) does not matter much in the error distribution (see the first two pairs of the boxplots), compared to when it is large (see the last two pairs). When the $|\lambda_0 - \lambda_1|$ is large, having constantly alternating Markov chain (or having $\det(A) \in [-1, -0.8]$) gives considerably more advantage to the MLE than

Table 1
The Root Mean Square Errors in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ with Various Combinations of $|\lambda_0 - \lambda_1|$ and $\det(A)$, $n = 30$

$\lvert\lambda_0 - \lambda_1\rvert$	$\det(A)$	mean		median		variance	
		MLE	BPM	MLE	BPM	MLE	BPM
$[0, 1]$	$[-1, -0.8]$	1.14	0.81	1.12	0.78	0.74	0.03
	$[0.8, 1]$	1.28	0.83	1.25	0.78	0.60	0.04
$(2, 3]$	$[-1, -0.8]$	0.42	1.35	0.35	1.43	0.29	0.16
	$[0.8, 1]$	1.28	1.57	1.02	1.67	0.95	0.42

having a Markov chain in which the current state does not have much effect on the probability distribution for the next state (or having $\det(A) \in [0.8, 1]$).

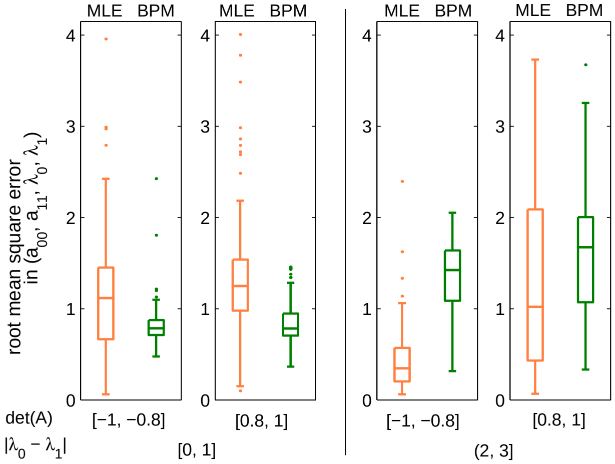


Fig. 5. The Root Mean Square Errors in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ with Various Combinations of $|\lambda_0 - \lambda_1|$ and $\det(A)$, $n = 30$

As for PHMMs with a smaller data size, the simulation implemented to produce Fig. 5 and Table 1 is repeated for sequence lengths $n = 10$ and 20 to see the effect of the data size to the error distribution. No statistically significant change is observed.

Also to see the overall errors, 1000 PHMMs are generated randomly with respect to $\det(A) \in [-1, 1]$ and $|\lambda_0 - \lambda_1| \in [0, 3]$ for each length $n = 10$, 20, and 30. The root mean square errors in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ for this simulation are plotted in Fig. 6. Again, no significant difference is observed in this data size range; but, BPM seems to have more stability than MLE in overall errors for PHMM parameter estimation at least when $n \leq 30$.

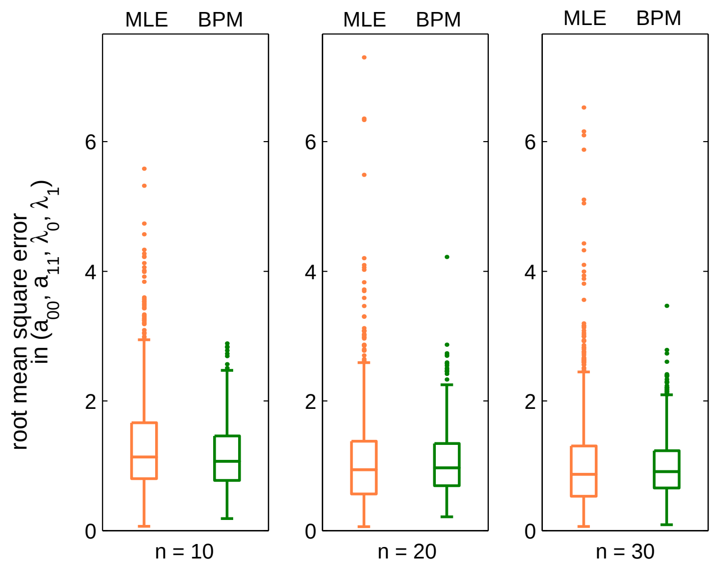


Fig. 6. The Overall Root Mean Square Errors in $(a_{00}, a_{11}, \lambda_0, \lambda_1)$ for $n = 10$, 20, and 30

Table 2
The Size of $\widehat{\Omega}_n$ Compared to the Size of Ω_n

n	ρ	size of Ω_n	size of $\widehat{\Omega}_n$					constant c
		(2^n)	mean	median	min.	max.	variance	
10	9	$1.02\ 10^3$	$2.72\ 10^2$	$2.80\ 10^2$	74	464	$7.5\ 10^3$	$4.46\ 10^{-23}$
20	12	$1.05\ 10^6$	$1.69\ 10^4$	$1.40\ 10^4$	344	$9.18\ 10^4$	$1.7\ 10^8$	$4.27\ 10^{-36}$
30	13	$1.07\ 10^9$	$2.72\ 10^5$	$1.77\ 10^5$	814	$2.03\ 10^6$	$8.6\ 10^{10}$	$3.66\ 10^{-43}$

Note: Constant c is such that maximum $= cn^{\nu(\nu+\rho)+3}$ with $\nu = 2$.

5.2 Computational complexity of the BPM

As a reference for the reduction in computational complexity for the BPM, simulation results are shown in Table 2, comparing the size of Ω_n, which is 2^n since $\nu = 2$ is used, and $\widehat{\Omega}_n$ with $n = 10, 20$, and 30. For each n, the data are for 1000 observation sequences generated from 1000 θ-values that are randomly picked so that they are uniformly distributed with respect to both $|\lambda_0 - \lambda_1|$ and $\det(A)$ under the restriction $0 < \lambda_0, \lambda_1 \leq 3$. The constant c is from the upper bound $cn^{\nu(\nu+\rho)+3} = cn^{2(2+\rho)+3}$ mentioned in Section 3.4.

6 Conclusions

Using rather simple but innovative algorithms described above, the exact BPM is actually feasible and is superior to the MLE for certain PHMM applications (and also for some other types of discrete HMM applications) if the application has small data size and observation space, but not so otherwise. How 'small' depends on today's computer capabilities and the actual implementation methods to realize the algorithm. Through simulations using the proposed algorithm applied to a two-state PHMM with small size data and observation space, we see how the MLE and BPM differ in their qualities. When the difference in the Poisson parameters is small, the BPM is closer to the true parameter set most frequently and also significantly more stable than the MLE, while when otherwise the MLE is closer to the true parameter set on average instead. At the same time, overall and on average the root mean square error is about the same, but the BPM is much more stable than the MLE. It would be worth considering the exact BPM, possibly together with the MLE (or with the approximation methods for BPM, like MCMC, for applications with the larger data size, using a part of the data, to get some extra intuition), when feasible.

Acknowledgments

I am deeply thankful to David Vere-Jones for the idea of extending the study on the BPM to PHMMs and also for his kind support in finishing up this paper. I also am very thankful to my Ph.D. adviser Tom Taylor; without his suggestions the research on the BPM never would have started. Furthermore, I would like to thank New Zealand Institute of Mathematics and its Applications (NZIMA) for its support.

A Integration Over the Transition Matrix A for BPM

As for the integration over A, given a Markov chain sequence of length n, $X^{1:n}$, we use the counts $K^{(n)} = \{k_{ij}^{(n)}\}_{i,j\in\zeta_X}$ obtained from $X^{1:n}$. We first denote an operator for addition in modulo as $\hat{+}$ so that

$$i\hat{+}j \equiv i + j \pmod{\nu} \quad \text{and} \quad i\hat{+}j \in \zeta_X.$$

Using this operator, we shift the entries of the transition matrix A within rows to have $\widetilde{A} = \{\tilde{a}_{ij}\}_{i,j\in\zeta_X}$, where $\tilde{a}_{ij}$ is defined as

$$\tilde{a}_{ij} = a_{i,\, i\hat{+}j}\,.$$

This is to make $\tilde{a}_{i0}$ the diagonal elements of A for notational convenience. Furthermore, define

$$\widetilde{K}^{(n)} = \{\tilde{k}_{ij}^{(n)}\}_{i,j\in\zeta_X} \quad \text{where} \quad \tilde{k}_{ij}^{(n)} = \tilde{k}_{ij}\left(X^{1:n}\right) = k_{i,\, i\hat{+}j}^{(n)}$$

so that the relationship between the counts $K^{(n)}$ and $\widetilde{K}^{(n)}$ is analogous to the one between A and $\widetilde{A}$. Then, in terms of $\widetilde{A}$ and $\widetilde{K}^{(n)}$, (5) can be written as

$$P(X^{1:n}, Y^{1:n} \mid \theta) = \pi_{x_1} \left(\prod_{i,j\,\in\,\zeta_X} \tilde{a}_{ij}^{\tilde{k}_{ij}\left(X^{1:n}\right)} \right) \left[\prod_{\substack{i\,\in\,\zeta_X \\ u\,\in\,\zeta_Y}} \left(f_{i,\lambda_i}(u)\right)^{l_{iu}\left(X^{1:n},\, Y^{1:n}\right)} \right].$$

For the integration over A to get $\Phi\left(K^{(n)}\right)$ and $\widetilde{\Phi}\left(K^{(n)}\right)$, we use only the product involving $\tilde{a}_{ij}$. We assumed Dirichlet distributions for the prior $P(A)$, while assuming row-wise independence of the elements in A, so that, for unknown parameters $\gamma_{i,0}, \gamma_{i,1}, \ldots, \gamma_{i,\nu-1}$, $i \in \zeta_X$,

$$P(A) = P(\widetilde{A}) = \prod_{i\,\in\,\zeta_X} P(\tilde{a}_{i0}, \tilde{a}_{i1}, \ldots \tilde{a}_{i,\nu-1}) = \prod_{i\,\in\,\zeta_X} \frac{\prod_{j\,\in\,\zeta_X} \tilde{a}_{ij}^{\gamma_{ij}-1}}{Z_i} = \prod_{i,j\,\in\,\zeta_X} \frac{\tilde{a}_{ij}^{\gamma_{ij}-1}}{Z_i}$$

where $Z_i = Z_i(\gamma_{i,0}, \ldots, \gamma_{i,\nu-1})$ are the normalizing constants such that the factor for $P(\tilde{a}_{i0}, \tilde{a}_{i1}, \ldots \tilde{a}_{i,\nu-1})$ integrates to unity for each i (Gelman et al., 1995).

Denote the range of $\widetilde{A}$ as $\mathcal{R}_{\widetilde{A}}$ so that $\widetilde{A}$ is a $\nu \times \nu$ probability matrix under the restriction $\tilde{a}_{i0} \geq \tilde{a}_{i+1,0}$ for all $i \in \{0,\, 1,\, \ldots,\, \nu-2\}$, which is equivalent to the original restriction $a_{ii} \geq a_{i+1,i+1}$.

From (7), since $\sum_{j\,\in\,\zeta_X} \tilde{a}_{ij} = \sum_{j=0}^{\nu-1} \tilde{a}_{ij} = 1$, we have

$$\begin{aligned}\Phi\left(K^{(n)}\right) &= \Phi_{D}\left(D\right) \\ &= \frac{1}{\prod_{i\,\in\,\zeta_X} Z_i} \int_{\widetilde{A}\in\mathcal{R}_{\widetilde{A}}} \prod_{i=0}^{\nu-1} \left[\left(1 - \sum_{j=0}^{\nu-2} \tilde{a}_{ij}\right)^{d_{i,\,\nu-1}} \prod_{j=0}^{\nu-2} \tilde{a}_{ij}^{d_{ij}} \right] \prod_{i=0}^{\nu-2} \prod_{j=0}^{\nu-2} d\tilde{a}_{ij},\end{aligned}$$

where

$$D = \{d_{ij}\}_{i,j\in\zeta_X},$$

and d_{ij} is the sum of the counts and the exponent of the prior defined as

$$d_{ij} = \tilde{k}_{ij}\left(X^{1:n}\right) + \gamma_{ij} - 1.$$

Then, by a straightforward (but rather long, owing to the restriction $\tilde{a}_{i0} \geq \tilde{a}_{i+1,0}$) derivation, we get

$$\begin{aligned} \Phi\left(K^{(n)}\right) &= \Phi_{_D}(D) \\ &= \frac{1}{\prod_{i \,\in\, \zeta_X} Z_i} \cdot G(D) \sum_{i_{\nu-1}=0}^{\tilde{p}_{\nu-1}} \sum_{i_{\nu-2}=0}^{\tilde{p}_{\nu-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} \left(\prod_{j=1}^{\nu-1} g_{j1}(I_j) \right) g_2(I_1), \end{aligned} \tag{A.1}$$

where

$$G(D) = \prod_{k \,\in\, \zeta_X} \left[\prod_{j\,=-1}^{\nu-4} \sum_{i=0}^{p_k(j)} g_{k0}(i,j) \right], \tag{A.2a}$$

$$g_{k0}(i,j) = \frac{(-1)^i p_k(j)!}{i!(p_k(j)-i)!(d_{k,\,\nu-j-3}+i+1)}, \tag{A.2b}$$

$$g_{k1}(I_k) = \frac{(-1)^{i_k} \tilde{p}_k!}{i_k!(\tilde{p}_k - i_k)!\phi_k(I_k)}, \text{ and} \tag{A.2c}$$

$$g_2(I_1) = \frac{(\phi_1(I_1) + d_{00})!\,\tilde{p}_0!}{(\phi_1(I_1) + d_{00} + \tilde{p}_0 + 1)!} \tag{A.2d}$$

for I_k, $p_k(i)$, $\tilde{p}_k$, and $\phi_k(I_k)$ defined as

$$I_k = (i_k, i_{k+1}, \ldots, i_{\nu-2}, i_{\nu-1}), \tag{A.2e}$$

$$p_k(i) = \sum_{j=\nu-i-2}^{\nu-1} d_{kj} + i + 1, \tag{A.2f}$$

$$\tilde{p}_k = \sum_{j=1}^{\nu-1} d_{kj} + \nu - 2 = p_k(\nu-3), \text{ and} \tag{A.2g}$$

$$\phi_k(I_k) = \sum_{j=k}^{\nu-1} (d_{j0} + i_j) + \nu - k. \tag{A.2h}$$

For the simulations in this paper, we let $\gamma_{ij} = 1$ for all $i, j \in \zeta_X$; i.e., we let $d_{ij} = \tilde{k}_{ij}$.

References

Albert, P.S., 1991. A two-state Markov mixture model for a time series of epileptic seizure counts. Biometrics, 47 1371-1381.

Baum, L.E., 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities, 3 1-8.

Baum, L.E., Petrie, T., Soules, G., & Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains. Ann. of Math. Statist., 41 164-171.

Capppé, O., Moulines, E., & Rydén, T., 2005. Inference in Hidden Markov Models. Springer-Verlag, New York.

Celeux, G., Hurn, M., & Robert, C.P., 2000. Computational and inferential difficulties with mixture posterior distributions. J. American Statist. Assoc., 95 957E70.

Ching, W.K., 1997. Markov-modulated Poisson processes for multi-location inventory problems. Internat. J. Production Econom., 53 217-223.

Chib, S., Chib. Calculating posterior distributions and modal estimates in Markov mixture models. J. Econom., 75 79-97.

Cooper, B., & Lipsitch, M., 2004. The analysis of hospital infection data using hidden Markov models. Biostatistics, 5(2) 223E37.

Dempster, A.D., Laird, N.M., & Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc., Ser. B, 39(1) 1-38.

Ephraim, Y., & Merhav, N., 2002. Hidden Markov processes. IEEE Transactions on Information Theory, 48(6) 1518-1569.

Gelman, A., Carlin, J., Stern, H., & Rubin, D.B., 1995. Bayesian Data Analysis. Springer-Verlag, New York.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice, Chapman & Hall, London.

Heffes, H., & Lucantoni, D.M., 1986. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. IEEE J. on Selected Areas in Comm., 4 856-868.

Heyman, D.P., & Lucantoni, D., 2003. Modeling multiple IP traffic streams with rate limits. IEEE/ACM Transactions on Networking, 11(6) 948-957.

Jana, R., & Dey, S., 2000. Change detection in teletraffic models. IEEE Trans. on Signal Processing, 48(3) 846-853.

MacDonald, I.L., & Zucchini, W., 1997. Hidden Markov and Other Models for Discrete-Valued Time Series. Chapman-Hall, London.

Murakami, J., & Taylor, T., 2006. Least square estimation of a hidden Markov chain parameters. Proc. of the 17th Internat. Sympos. on Math. Theory of Networks and Systems (MTNS), July 2006, Kyoto, Japan, 2151-2156.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 77(2) 257-286.

Robert, C.P., & Titterington, D.M., 1998. Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. Stat. Computing, 8 145-158.

Scott, S.L., 2002. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. J. Amer. Statist. Assoc., 97 337-351.

Scott, S.L., & Smyth, P., 2003. The Markov Modulated Poisson Process and Markov Poisson Cascade with Applications to Web Traffic Modeling. Bayesian Statist., 7 671-680, Oxford University Press.

Stephens, M., 2000. Dealing with label switching in mixture models. J. R. Stat. Soc. Ser. B Stat. Methodol., 62(4) 795E09.