PARAMETER ESTIMATE OF A HIDDEN MARKOV CHAIN

by

Junko Murakami

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

May 2005

ABSTRACT

Hidden Markov Models (HMM) are of considerable interest for science and for various applications. They consist of a Markov chain with "hidden" states and emissions which are statistically dependent on the states but can be observed. The model is parameterized by two conditional probability matrices, the transition and emission matrices.

Among the algorithms used for the parameter estimate of HMM, Baum-Welch (B-W) algorithm is by far the most popular algorithm. However, it has some well-known shortcomings. For example, it is only guaranteed to find a local maximum, with a strong dependency on the initial parameters chosen. The literature notes an "overfitting" problem, in which the B-W estimate gives high likelihood to a given observation sequence, but low likelihood to other observation sequences of the same hidden Markov chain. As a consequence of these this researcher has shown that usually the B-W estimator is inconsistent for the simplest possible case of two hidden states and two emission states, and with a generic choice of parameters and generic observation sequences.

The dissertation also provides an algorithm for computation of the least square error (LSE) estimate of the hidden Markov chain. The LSE estimate is consistent by definition, needs only one time computation, and again with the simplest possible case of HMM described above this researcher has demonstrated the estimates are remarkably closer to the actual parameters, with much better results than what could typically be obtained using the B-W algorithm. A possible reason for the LSE estimation not being very popular regarding HMM, in spite of its much superior quality of its estimates, could be the computational complexity it requires. Although a straightforward computation could require the complexity that increases exponentially with respect to the sequence length, this researcher has shown that a polynomial complexity (with still exponential complexity in the

state size) can be achieved using an algorithm proposed in this dissertation, making the LSE estimation quite feasible in some applications such as the ones related to precipitation, heart rate monitoring, and so on.

To my mother and my late father, and to the memory of Yumiko Hasegawa

ACKNOWLEDGMENTS

I wish to thank my thesis adviser Dr. Thomas Taylor for his great ideas and guidance, without which this dissertation would not have been possible, and Dr. Roger Brockett for the suggestion to investigate the dynamics of the Baum-Welch algorithm.

I also wish to thank my mother for her never-ending support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

CHAPTER 1

Introduction

In an HMM, an underlying unobservable Markov chain emits an output sequence that can be observed. This model can be applied in various fields, and so the Baum-Welch (B-W) algorithm [3, 4], which is one implementation of the Expectation Maximization (EM) algorithm [8, 28], is used in numerous applications. Most often the algorithm is used in speech recognition [9, 14, 15, 16, 18, 21, 27, 30]; but, also widely used in various other fields. For example, it is used in biology [11], especially in gene analysis [**?**, 26, 32]; human activity analysis [24]; data prefetching [17]; political event prediction [29]; visual pattern recognition [7], which includes handwritten text analysis [13], facial expression recognition [22, 6], and action recognition [19]; magnetic recording channel analysis [1]; estimation of packet loss on internet path [31]; precipitation models [33]; heart rate variability [12]; and so on.

However, a well-known limitation of the E-M algorithm are that it is only guaranteed to find a local optimum which, under a certain conditions, strongly depends on the initial estimate [8, 23]. It also has an "overfitting" problem. Specifically, it fits the output sequence well but the HMM which generated the sequence poorly. On the other hand, the least square error (LSE) estimator can be obtained using a deterministic formula. However, the computational complexity increases exponentially with respect to the length of the observation sequence if it is obtained in a naive way. The algorithm that we introduce here

reduces the computational complexity to polynomial with respect the sequence length. The complexity is still exponential in the state space size, however.

Also, the researcher studied the convergence of the B-W algorithm. The iterative re-estimation of the parameter is a non-linear mapping, which stops when it converges to a fixed point. By experimentally finding the Jacobian matrices for the mapping, which allows us to see the linearization of the algorithm's behavior, we studied the rate of convergence.

The B-W algorithm is a special case of EM algorithm. In the case of Gaussian mixtures, Ma, Xu, and Jordan [20] had established super-exponential convergence in some cases. In contrast of this, our empirical results show that numerical B-W "fixed points" have linear convergence at best. It is also revealed that a certain number of numerical "fixed points" are likely to approximate saddle points of the likelihood function, rather than maxima.

In Chapter 2 the B-W algorithm is described with the HMM in which both state space size and emission state size are two. Then the formula for the Jacobian is obtained, which is used to empirically find the Jacobian determinant of the B-W algorithm for such HMM. In Chapter 3, for the HMM with the state space size m, where $m \geq 2$ is any positive integer, and the emission space size 2, the exact formula for the LSE estimation is shown, together with the algorithm that significantly reduces the computational complexity with respect to the sequence length. The formula and the algorithm are then applied to the cases of state space size two, three, and five. Furthermore, the particle filter method is described, which can be used to approximate the LSE estimation. In Chapter 4, experimental results are summarized, first for the B-W algorithm, then a comparison is made with LSE estimation mainly regarding the "overfitting" problem. The empirical results show that the most of the time the B-W algorithm maximizes the likelihood with HMMs with its state

and emission state size two; but, it does have the "overfitting" problem, while the LSE estimator has no such problem and it is closer to the true parameter than the B-W estimator. The above conclusions are summarized in Chapter 5.

1. Hidden Markov Models

In a hidden Markov model (HMM), we have two sequences of states for discrete time t: one for the states of a Markov chain (the state sequence) and one for the observed states (the observation sequence). The states of the Markov chain cannot be observed (are hidden), while the observed states depend on them.

Let $S_1, S_2, S_3, \ldots$ be a Markov chain, and assume we only know the corresponding observed states $O_1, O_2, O_3, \ldots$ Although the dimensions of the state space are mainly set to two for both sequences in this paper, they can be easily extended to a larger size.

Denote state sequences as $S^{m,n} = (S_m, S_{m+1}, \ldots, S_n)$ and observation sequences as $O^{m,n} = (O_m, O_{m+1}, \ldots, O_n)$. With a HMM, we have the following properties:

$$P(S_{t+1}|O^{1,t}, S^{1,t}) = P(S_{t+1}|S_t) \tag{1.1}$$

and

$$P(O_t|O^{1,t-1}, O^{t+1,n}, S^{1,n}) = P(O_t|S_t) \tag{1.2}$$

for any $1 \leq t \leq n$.

The first property (1.1) is a property of a Markov chain: The probability of transition from a state S_t at time t to the next state S_{t+1} at time $t+1$ depends only on the current state S_t, and not on any prior states. The second property (1.2) shows that the probability

of the observed state O_t at time t depends only on the current state of the Markov chain state S_t, and not on any other states of the state sequence nor on any other observed states.

Hence, to describe a HMM, we need two probability matrices – one for the state sequence (of the Markov chain) and the other for the observation sequence – plus a probability vector for the initial states. Let the state space sizes for the state sequence and the observation sequence be m_A and m_B, respectively. Also, for simplicity, we assign non-negative integers to the states. Let the state space be $\{0, 1, 2, \ldots, m_A - 1\}$ and $\{0, 1, 2, \ldots, m_B - 1\}$ for the state sequence and the observation sequence, respectively. Furthermore, let the matrix for the state sequence be $A = \{a_{ij}\}$, the matrix for the observation sequence be $B = \{b_{ik}\}$, and the vector of probabilities for the initial state be $\pi = (\pi_0, \pi_1, \ldots, \pi_{m_A-1})$, where $i, j \in \{0, 1, \ldots, m_A - 1\}$ and $k \in \{0, 1, \ldots, m_B - 1\}$ so that, for any t,

$$a_{ij} = P(S_{t+1} = j \mid S_t = i),$$

$$b_{ik} = P(O_t = k \mid S_t = i), \text{ and}$$

$$\pi_i = P(S_1 = i).$$

Hence, A is a $m_A \times m_A$ matrix, B is a $m_A \times m_B$ matrix, and π is a vector of length m_A.

Now, we let $\theta = \{\pi, A, B\}$, the parameter set for the HMM.

For example, consider the case $m_A = m_B = 2$; i.e., the state space is $\{0, 1\}$ for both sequences, A and B are 2×2 matrices, and π is a length 2 vector, and the HMM can be described as follows:

Let

$$A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} = \begin{pmatrix} a & a-1 \\ b-1 & b \end{pmatrix} \tag{1.3}$$

be the transition matrix (for the state sequence) and

$$B = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix} = \begin{pmatrix} x & 1-x \\ 1-y & y \end{pmatrix} \tag{1.4}$$

be the emission matrix (for the observation sequence) so that, for all discrete time $t \geq 1$ and i, j, $k \in \{0, 1\}$,

$$a_{ij} = P(S_{t+1} = j \mid S_t = i)$$

is the probability of transferring from state i to state j, and

$$b_{ik} = P(O_t = k \mid S_t = i)$$

is the probability of observing state k when the transition state is i. Also, let

$$\pi_i = P(S_1 = i), \qquad i = 0, 1,$$

be the probability for the initial state of the state sequence being i, and let

$$\pi = (\pi_0, \pi_1)^T = (r, 1-r)^T. \tag{1.5}$$

We try to find an estimator for $\theta = \{\pi, A, B\}$, and in this paper we consider two types of parameter estimators, which are the maximum likelihood estimator (using the B-W algorithm) and the least square estimator. In general, these two estimation methods can be described as follows:

Let $X = (X_1, X_2, \ldots, X_n)$ be random variables with the probability distribution function $f(x \mid \theta)$, where $x = x_1, x_2, \ldots, x_n)$ is a particular outcome and θ is a set of parameters to estimate.

2. Maximum Likelihood Estimation

We find the value for θ so that the probability of a particular outcome x is as high as possible.

The function $L_x(\theta) = f(x \mid \theta)$, the probability of having a particular outcome x given the parameter being θ, is called a likelihood function. So, the estimator, θ_{ML}, can be expressed as

$$\theta_{ML} = arg \max_{\theta} L_x(\theta).$$

For the same purpose, instead of the likelihood function above, the log likelihood function $\log L_x(\theta)$ can be used, which gives

$$\theta_{ML} = arg \max_{\theta} [\log L_x(\theta)].$$

3. Least Square Estimation

Let $\theta' = \theta'(X)$ be the estimator of θ. We find θ' so that the mean square error of θ', $E\|\theta'(x) - \theta\|^2$, is as small as possible. Then, $\theta'(x)$ is $E(\theta \mid x)$, the expected value of the parameter set given particular observed data x. Thus, the estimator θ_{LS} can be expressed as

$$\theta_{LS} = E(\theta \mid x).$$

By the definition, the estimate is unbiased and the variance is minimum.

CHAPTER 2

Baum-Welch Algorithm

1. What is the Baum-Welch Algorithm?

Let Ω_n be the space of all possible state sequences $S^{1,n} = (S_1, S_2, \ldots, S_n)$, $S_i \in \{0, 1\}$, $i = 1, \ldots, n$. As before, let $O^{1,n} = (O_1, O_2, \ldots, O_n)$ be the observation sequence. The goal is to find a parameter set θ that maximizes the likelihood function $L(\theta) = P(O^{1,n}, S^{1,n} \mid \theta)$. So, we try to maximize the expected value of its log likelihood function.

As one of the implementations of EM Algorithm, B-W algorithm consists of E-steps and M-steps.

E-step: Let $\hat{\theta}$ be the current estimate, and define a function Q for the expected value of the log likelihood function $L(\theta) = P\left(O^{1,n}, S^{1,n} \mid \theta\right)$ (the probability of sequences $O^{1,n}$ and $S^{1,n}$ given that the parameter is θ) given an observation sequence $O^{1,n}$ and the current parameter estimate $\hat{\theta}$ as shown below.

$$\begin{aligned} Q(\theta, \hat{\theta}) &= E\left[\log L(\theta) \mid O^{1,n}, \hat{\theta}\right] \\ &= E\left[\log P\left(O^{1,n}, S^{1,n} \mid O^{1,n}, \theta\right) \mid O^{1,n}, \hat{\theta}\right] \\ &= \sum_{s^{1,n} \in \Omega_n} \left[\log P\left(O^{1,n}, s^{1,n} \mid \theta\right)\right] P\left(s^{1,n} \mid O^{1,n}, \hat{\theta}\right). \end{aligned}$$

We are supposed to look for the value of θ that will maximize Q. However, $P\left(s \mid O^{1,n}, \hat{\theta}\right)$

is hard to find. But, using the Baye's formula, we get

$$P\left(s \mid O^{1,n}, \hat{\theta}\right) = \frac{P\left(O^{1,n}, s \mid \hat{\theta}\right)}{P\left(O^{1,n} \mid \hat{\theta}\right)}.$$

We note that $P\left(O^{1,n} \mid \hat{\theta}\right)$ does not depend on θ, so that in place of $\arg\max_\theta Q$ it is easier to compute $\arg\max_\theta \widetilde{Q}$ where

$$\widetilde{Q}\left(\theta, \hat{\theta}\right) = \sum_{s^{1,n} \in \Omega_n} \log\left[P\left(O^{1,n}, s^{1,n} \mid \theta\right)\right] P\left(O^{1,n}, s^{1,n} \mid \hat{\theta}\right).$$

Note that $s^{1,n}$ is a random variable here, while $O^{1,n}$ and $\hat{\theta}$ are constants.

M-Step: As for the iteration steps, what we have is then

$$\theta^{(k+1)} = \arg\max_\theta \widetilde{Q}\left(\theta, \theta^{(k)}\right), k = 0, 1, 2, \ldots$$

For notational convenience, let us consider $S^{0,n} = (S_0, S_1, S_2, \ldots, S_n)$ instead of $S^{1,n}$. Then we can write, for particular sequences $s^{0,n}$ and $o^{1,n}$,

$$P\left(o^{1,n}, s^{0,n} \mid \theta\right) = \pi_{s_0}\left(a_{s_0 s_1} b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n}\right) = \pi_{s_0} \prod_{t=1}^{n} a_{s_{i-1} s_i} b_{s_i o_i}.$$

Taking the log of the above, we can use the Lagrange multiplier to optimize $\widetilde{Q}$ term-wise individually for each parameter. The result is the Baum-Welch algorithm [28, 3, 5].

1.1. Derivation. First, we rewrite Q so that we can maximize term-wise for each parameters.

$$\widetilde{Q}\left(\theta, \hat{\theta}\right) = \sum_{s^{1,n} \in \Omega_n} \left[P\left(O^{1,n}, s^{1,n} \mid \hat{\theta}\right) \log \pi_{s_0} + \sum_{t=1}^{n} P\left(O^{1,n}, s^{1,n} \mid \hat{\theta}\right) \log a_{s_t - 1 s_t} + \sum_{t=1}^{n} P\left(O^{1,n}, s^{1,n} \mid \hat{\theta}\right) \log b_{s_t O_t} \right]$$

Now the first term is actually

$$\sum_{s^{1,n}\in\Omega_n} P\left(O^{1,n}, s^{1,n} \mid \hat{\theta}\right) \log \pi_{s_0} = \sum_{i=0}^{1} P\left(O^{1,n}, s_0 = i \mid \hat{\theta}\right) \log \pi_i$$

because by summing up all the possible state sequence values, we end up having the marginal distribution for the first state. Using the constraint $\sum_{i=0}^{1} \pi_i = 1$ and the Lagrange multiplier λ, and letting the partial derivative be zero, we get, for $i = 0, 1$,

$$\frac{\partial}{\partial \pi_i}\left[\sum_{j=0}^{1} P\left(O^{1,n}, s_0 = j \mid \hat{\theta}\right) \log \pi_i + \lambda\left(\sum_{j=0}^{1} \pi_j - 1\right)\right] = 0$$

$$\frac{1}{\pi_i} P\left(O^{1,n}, s_0 = i \mid \hat{\theta}\right) + \lambda = 0$$

so that

$$\frac{1}{\pi_0} P\left(O^{1,n}, s_0 = 0 \mid \hat{\theta}\right) = \frac{1}{\pi_1} P\left(O^{1,n}, s_0 = 1 \mid \hat{\theta}\right).$$

Solving the above, we get the formula for the next estimate of π_i as below.

$$\pi_i = \frac{P\left(O^{1,n}, s_0 = i \mid \hat{\theta}\right)}{\sum_{j=0}^{1} P\left(O^{1,n}, s_0 = j \mid \hat{\theta}\right)} = \frac{P\left(O^{1,n}, s_0 = i \mid \hat{\theta}\right)}{P\left(O^{1,n} \mid \hat{\theta}\right)} \tag{2.1}$$

The second term involving a_{ij} can be rewritten as

$$\sum_{s^{1,n}\in\Omega_n} \sum_{t=1}^{n} P\left(O^{1,n}, s^{1,n} \mid \hat{\theta}\right) \log a_{s_{t-1}s_t} = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{t=1}^{n} P\left(O^{1,n}, s_{t-1} = i, s_t = j \mid \hat{\theta}\right)$$

where the right hand side is the marginal probability for $t-1$ and t. So, using the restriction $\sum_{j=0}^{1} a_{ij} = 1$ for $i = 0, 1$ and the Lagrange multipliers λ_0 and λ_1, we have

$$\frac{\partial}{\partial a_{ij}}\left[\sum_{k=0}^{1}\sum_{l=0}^{1}\sum_{t=1}^{n} P\left(O^{1,n}, s_{t-1} = k, s_t = l \mid \hat{\theta}\right) \log a_{kl} + \sum_{k=0}^{1} \lambda_k \left(\sum_{l=0}^{1} a_{kl} - 1\right)\right] = 0.$$

Taking the partial derivatives, we get

$$\frac{1}{a_{ij}} \sum_{t=1}^{n} P\left(O^{1,n}, s_{t-1} = i, s_t = j \mid \hat{\theta}\right) + \lambda_i = 0 \qquad \text{for } i, j \in \{0, 1\}.$$

Solving this system of four equations, we see that the next estimate for a_{ij} should be set as

$$a_{ij} = \frac{\sum_{t=1}^{n} P\left(O^{1,n}, s_{t-1} = i, s_t = j \mid \hat{\theta}\right)}{\sum_{k=0}^{1}\sum_{t=1}^{n} P\left(O^{1,n}, s_{t-1} = i, s_t = k \mid \hat{\theta}\right)} = \frac{\sum_{t=1}^{n} P\left(O^{1,n}, s_{t-1} = i, s_t = j \mid \hat{\theta}\right)}{\sum_{t=1}^{n} P\left(O^{1,n}, s_{t-1} = i \mid \hat{\theta}\right)} \tag{2.2}$$

Finally as for the parameters b_{ij}, we can rewrite the third term as

$$\sum_{s^{1,n}\in\Omega_n}\sum_{t=1}^{n} P\left(O^{1,n}, s^{1,n} \mid \hat{\theta}\right)\log b_{s_t O_t} = \sum_{i=0}^{1}\sum_{t=1}^{n} P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)\log b_{iO_t}$$

where the right hand side is the marginal distribution for time t. Then, using the restriction $\sum_{j=0}^{1} b_{ij} = 1$ for $i = 0, 1$ and the Lagrange multipliers λ_0 and λ_1, we have

$$\frac{\partial}{\partial b_{ij}}\left[\sum_{k=0}^{1}\sum_{t=1}^{n} P\left(O^{1,n}, s_t = k \mid \hat{\theta}\right)\log b_{kO_t} + \sum_{k=0}^{1}\lambda_k\left(\sum_{l=0}^{1} b_{kl} - 1\right)\right] = 0.$$

Since only the terms with b_{ij} counts when taken the partial derivative, this implies

$$\frac{\partial}{\partial b_{ij}}\left[\sum_{t=1}^{n} P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)\log b_{iO_t} + \sum_{k=0}^{1}\lambda_k\left(\sum_{l=0}^{1} b_{kl} - 1\right)\right] = 0$$

so that we have

$$\frac{1}{b_{ij}}\sum_{t=1}^{n} P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)\delta_j(O_t) + \lambda_i = 0 \qquad \text{for } i = 0, 1$$

where δ is defined as

$$\delta_j(v) = \begin{cases} 1 & \text{if } v = j \\ 0 & \text{if } v \neq j\,. \end{cases}$$

Therefore, we set the next estimates for b_{ij} to be

$$b_{ij} = \frac{\sum_{t=1}^{n} P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)\delta_j(O_t)}{\sum_{k=0}^{1}\sum_{t=1}^{n} P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)\delta_k(O_t)} = \frac{\sum_{t=1}^{n} P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)\delta_j(O_t)}{\sum_{t=1}^{n} P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)}. \tag{2.3}$$

Now, in order to simplify the expressions for the next estimates of π_i, a_{ij}, and b_{ij} that we have just obtained above, i.e., (2.1), (2.2), and (2.3), we define $\gamma_i(t)$ and $\xi_{ij}(t)$ for

$i, j \in \{0, 1\}$ as follows:

$$\gamma_i(t) = P\left(s_t = i \mid O^{1,n}, \hat{\theta}\right) = \frac{P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)}{P\left(O^{1,n} \mid \hat{\theta}\right)} \tag{2.4}$$

$$\xi_{ij}(t) = P\left(s_t = i, s_{t+1} = j \mid O^{1,n}, \hat{\theta}\right) = \frac{P\left(s_t = i, s_{t+1} = j \mid O^{1,n}, \hat{\theta}\right)}{P\left(O^{1,n} \mid \hat{\theta}\right)} \tag{2.5}$$

Then, the estimates can be rewritten using $\gamma_i(t)$ and $\xi_{ij}(t)$ as shown below.

$$\pi_i = \gamma_i(t) \tag{2.6}$$

$$a_{ij} = \frac{\sum_{t=1}^{n-1} \xi_{ij}(t)}{\sum_{t=1}^{n-1} \gamma_i(t)} \tag{2.7}$$

$$b_{ij} = \frac{\sum_{t=1}^{n} \delta_j(O_t)\gamma_i(t)}{\sum_{t=1}^{n} \gamma_i(t)} \tag{2.8}$$

For actual computation, we further define two more variables, $\alpha_i(t)$ and $\beta_i(t)$ for $i = 0, 1$, as

$$\alpha_i(t) = P\left(O^{1,t}, s_t = i \mid \hat{\theta}\right) \qquad \text{and} \qquad \beta_i(t) = P\left(O^{t+1,n} \mid s_t = i, \hat{\theta}\right), \tag{2.9}$$

and express $\gamma_i(t)$ and $\xi_{ij}(t)$ using these variables. Basically, in order to do so, we just need to use the probability property $P(A, B) = P(A \mid B)P(B)$ and the properties of HMM to rewrite $\gamma_i(t)$ and $\xi_{ij}(t)$ as shown below. Given the current parameter set $\hat{\theta}$, we have

$$\begin{aligned} P\left(O^{1,n}, s_t = i\right) &= P\left(O^{1,t}, O^{t+1,n}, s_t = i\right) \\ &= P\left(O^{1,t}, s_t = i\right) P\left(O^{t+1,n} \mid O^{1,t}, s_t = i\right) \\ &= P\left(O^{1,t}, s_t = i\right) P\left(O^{t+1,n} \mid s_t = i\right) \\ &= \alpha_i(t)\beta_i(t) \end{aligned}$$

which implies

$$\gamma_i(t) = \frac{P\left(O^{1,n}, s_t = i \mid \hat{\theta}\right)}{P\left(O^{1,n} \mid \hat{\theta}\right)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{v=0}^{1} \alpha_v(t)\beta_v(t)}$$

and

$$\begin{aligned}
P\left(O^{1,n}, s_t = i, s_{t+1} = j\right) &= P\left(O^{1,t}, O_{t+1}, O^{t+2,n}, s_t = i, s_{t+1} = j\right) \\
&= P\left(O^{1,t}, s_t = i\right) P\left(O_{t+1}, O^{t+2,n}, s_{t+1} = j \mid O^{1,t}, s_t = i\right) \\
&= P\left(O^{1,t}, s_t = i\right) P\left(s_{t+1} = j \mid O^{1,t}, s_t = i\right) \\
&\qquad P\left(O_{t+1}, O^{t+2,n} \mid O^{1,t}, s_t = i, s_{t+1} = j\right) \\
&= P\left(O^{1,t}, s_t = i\right) P\left(s_{t+1} = j \mid s_t = i\right) \\
&\qquad P\left(O_{t+1} \mid O^{1,t}, s_t = i, s_{t+1} = j\right) \\
&\qquad P\left(O^{t+2,n} \mid O^{1,t}, O_{t+1}, s_t = i, s_{t+1} = j\right) \\
&= P\left(O^{1,t}, s_t = i\right) P\left(s_{t+1} = j \mid s_t = i\right) \\
&\qquad P\left(O_{t+1} \mid s_{t+1} = j\right) P\left(O^{t+2,n} \mid s_{t+1} = j\right) \\
&= \alpha_i(t)\, \hat{a}_{ij}\, \hat{b}_{j\, O_{t+1}}\, \beta_j(t+1)
\end{aligned}$$

which implies

$$\xi_{ij}(t) = \frac{P\left(O^{1,n}, s_t = i, s_{t+1} = j \mid \hat{\theta}\right)}{P\left(O^{1,n} \mid \hat{\theta}\right)} = \frac{\alpha_i(t)\, \hat{a}_{ij}\, \hat{b}_{j\, O_{t+1}}\, \beta_j(t+1)}{\sum_{v=0}^{1} \sum_{w=0}^{1} \alpha_v(t)\, \hat{a}_{vw}\, \hat{b}_{w\, O_{t+1}}\, \beta_w(t+1)}$$

where $\hat{\theta} = \{\hat{\pi}, \hat{A}, \hat{B}\}$, $\hat{A} = \{\hat{a}_{ij}\}$, and $\hat{B} = \{\hat{b}_{ij}\}$. So, we see that if the values of $\alpha_i(t)$ and $\beta_i(t)$ are obtained, then we can find the values of $\gamma_i(t)$ and $\xi_{ij}(t)$, which then give the value of the next estimate $\theta^{(k+1)}$ can be found given the current estimate $\hat{\theta} = \theta^{(k)}$. But, $\alpha_i(t)$ and $\beta_i(t)$ can be found recursively as we can see by rewriting them using the HMM properties (1.1) and (1.2), and using the concept of marginal distributions.

First, we express $\alpha_i(t+1)$ using $\alpha_i(t)$ as follows:

By the definition (2.9), given $\hat{\theta}$, $\alpha_i(t)$ can be expressed as

$$\begin{aligned}\alpha_i(t) &= P\left(O^{1,t}, s_t = i\right) \\ &= \sum_{S^{1,t-1} \in \Omega_{t-1}} P\left(O^{1,t}, S^{1,t-1}, S_t = i\right) \\ &= \sum_{S^{1,t-1} \in \Omega_{t-1}} P\left(O^{1,t} \mid S^{1,t-1}, S_t = i\right) P\left(S^{1,t-1}, S_t = i\right),\end{aligned}$$

and so $\alpha_i(t+1)$ is

$$\alpha_i(t+1) = \sum_{S^{1,t} \in \Omega_t} P\left(O^{1,t+1} \mid S^{1,t}, S_{t+1} = i\right) P\left(S^{1,t}, S_{t+1} = i\right).$$

Then

$$\begin{aligned}\alpha_i(t+1) &= \sum_{S^{1,t} \in \Omega_t} P\left(O^{1,t}, O_{t+1} \mid S^{1,t}, S_{t+1} = i\right) P\left(S_{t+1} = i \mid S^{1,t}\right) P\left(S^{1,t}\right) \\ &= \sum_{S^{1,t} \in \Omega_t} P\left(O^{1,t} \mid S^{1,t}\right) P\left(O_{t+1} \mid S_{t+1} = i\right) P\left(S_{t+1} = i \mid S_t\right) P\left(S^{1,t}\right) \\ &= \sum_{S^{1,t} \in \Omega_t} P\left(O^{1,t} \mid S^{1,t-1}, S_t\right) P\left(O_{t+1} \mid S_{t+1} = i\right) P\left(S_{t+1} = i \mid S_t\right) P\left(S^{1,t-1}, S_t\right) \\ &= \sum_{j \in \{0,1\}} P\left(S_{t+1} = i \mid S_t = j\right) \sum_{S^{1,t-1} \in \Omega_{t-1}} P\left(O^{1,t} \mid S^{1,t-1}, S_t = j\right) P\left(S^{1,t-1}, S_t = j\right) \\ &\qquad \cdot P\left(O_{t+1} = o_{t+1} \mid S_{t+1} = i\right) \\ &= \left(\sum_{j \in \{0,1\}} a_{ji}\, \alpha_j(t)\right) b_{io_{t+1}}.\end{aligned}$$

Similarly, we can express $\beta_i(t)$ using $\beta_i(t+1)$ as follows:

Given $\hat{\theta}$, we have

$$\begin{aligned}
\beta_i(t+1) &= P\left(O^{t+2,n} \mid S_{t+1} = i\right) \\
&= \sum_{S^{t+2,n} \in \Omega_{n-t-1}} P\left(O^{t+2,n}, S^{t+2,n} \mid S_{t+1} = i\right) \\
&= \sum_{S^{t+2,n} \in \Omega_{n-t-1}} P\left(O^{t+2,n} \mid S^{t+2,n}, S_{t+1} = i\right) P\left(S^{t+2,n} \mid S_{t+1} = i\right) \\
&= \sum_{S^{t+2,n} \in \Omega_{n-t-1}} P\left(O^{t+2,n} \mid S^{t+2,n}\right) P\left(S^{t+2,n} \mid S_{t+1} = i\right)
\end{aligned}$$

and so

$$\beta_i(t) = \sum_{S^{t+1,n} \in \Omega_{n-t}} P\left(O^{t+1,n} \mid S^{t+1,n}\right) P\left(S^{t+1,n} \mid S_t = i\right).$$

Then, we get

$$\begin{aligned}
\beta_i(t) &= \sum_{S^{t+1,n} \in \Omega_{n-t}} P\left(O_{t+1}, O^{t+2,n} \mid S_{t+1}, S^{t+2,n}\right) P\left(S_{t+1}, S^{t+2,n} \mid S_t = i\right) \\
&= \sum_{S^{t+1,n} \in \Omega_{n-t}} P\left(O_{t+1} \mid O^{t+2,n}, S_{t+1}, S^{t+2,n}\right) P\left(O^{t+2,n} \mid S_{t+1}, S^{t+2,n}\right) \\
&\qquad \cdot P\left(S^{t+2,n} \mid S_{t+1}, S_t = i\right) P\left(S_{t+1} \mid S_t = i\right) \\
&= \sum_{S^{t+1,n} \in \Omega_{n-t}} P\left(O_{t+1} \mid S_{t+1}\right) P\left(O^{t+2,n} \mid S^{t+2,n}\right) \\
&\qquad \cdot P\left(S^{t+2,n} \mid S_{t+1}\right) P\left(S_{t+1} \mid S_t = i\right) \\
&= \sum_{j \in \{0,1\}} P\left(S_{t+1} = j \mid S_t = i\right) P\left(O_{t+1} = o_{t+1} \mid S_{t+1} = j\right) \\
&\qquad \cdot \sum_{S^{t+2,n} \in \Omega_{n-t-1}} P\left(O^{t+2,n} \mid S^{t+2,n}\right) P\left(S^{t+2,n} \mid S_{t+1} = j\right) \\
&= \sum_{j \in \{0,1\}} a_{ij}\, b_{j\, o_{t+1}} \beta_j(t+1).
\end{aligned}$$

Thus, we get the B-W algorithm, which is outlined and then described below.

1.2. Baum-Welch Algorithm. Let $\theta^{(k)} = \left(\pi^{(k)}, A^{(k)}, B^{(k)}\right)$ be the kth estimate of θ.

For $k = 0$: Choose the initial parameter estimate $\theta^{(k)} = \theta^{(0)}$.

For $k > 0$: Repeat until $\theta^{(k)} \approx \theta^{(k+1)}$.

Using the current estimate $\theta^{(k)}$

1. forward procedure: Find $\alpha_i^{(k)}(t)$, which is the probability of S_t being i and the partial observation sequence up to time t being $o^{1,t} = (o_1, o_2, \ldots, o_t)$, for $t = 1$ to n, $i = 0, 1$.

2. backward procedure: Find $\beta_i^{(k)}(t)$, which is the probability of partial observation sequence after time t being $o^{t+1,n} = (o_{t+1}, o_{t+2}, \ldots, o_n)$, given that S_t being i, for $t = n$ down to 1, $t = 0, 1$.

3. Using the values of $\alpha_i^{(k)}(t)$ and $\beta_i^{(k)}(t)$, find $\gamma_i^{(k)}(t)$, which is the probability of S_t being i, for $t = 1$ to n, $i = 0, 1$.

4. Using the values of $\gamma_i^{(k)}(t)$, $\beta_i^{(k)}(t)$, and $\beta_i^{(k)}(t+1)$, find $\xi_{ij}^{(k)}(t)$, which is the probability of the current state S_t being i and the next state S_{t+1} being j, for $t = 1$ to $n - 1$, $i, j = 0, 1$.

5. Find a new estimate $\theta^{(k+1)}$, using the fact that $\sum_{t=1}^{n} \gamma_i^{(k)}(t)$ is the expected total number of transitions away from state i, and $\sum_{i=1}^{n-1} \xi_{ij}^{(k)}(t)$ is the expected number of transitions from state i to state j in the sequence $S^{0,n}$.

A little more detailed algorithm for case the state space for both state and observation sequences being $\{0, 1\}$ is as follows:

MAIN (The Baum-Welch Algorithm)

repeat until $\theta^{(k)} \approx \theta^{(k+1)}$

$\alpha^{(k)} = \text{forward\_procedure } \left(O^{1,n}, \theta^{(k)}\right)$

$\beta^{(k)} = \text{backward\_procedure } \left(O^{1,n}, \theta^{(k)}\right)$

$\gamma^{(k)} = \text{get\_}\gamma \left(\alpha^{(k)}, \beta^{(k)}\right)$

$\xi^{(k)} = \text{get\_}\xi \left(O^{1,n}, \alpha^{(k)}, \beta^{(k)}, \theta^{(k)}\right)$

$\theta^{(k+1)} = \text{get\_new\_estimates } \left(O^{1,n}, \gamma^{(k)}, \xi^{(k)}\right)$

end

---

forward_procedure $\left(O^{1,n}, \theta\right)$

$\alpha_i(1) = \pi_i b_{i,o_1}$ for $i = 0, 1$

$\alpha_i(t+1) = \left(\alpha_0(t)a_{0i} + \alpha_1(t)a_{1i}\right) b_{i,o_{t+1}}$ for $i = 0, 1$ and $t = 1, \ldots, n-1$

$$\text{return } \alpha = \begin{pmatrix} \alpha_0(1) & \cdots & \alpha_0(n) \\ \alpha_1(1) & \cdots & \alpha_1(n) \end{pmatrix}$$

---

backward_procedure $\left(O^{1,n}, \theta\right)$

$\beta_i(n) = 1$ for $i = 0, 1$

$\beta_i(t) = a_{i0}b_{0,o_{t+1}}\beta_0(t+1) + a_{i1}b_{1,o_{t+1}}\beta_1(t+1)$ for $i = 0, 1$ and $t = 1, \ldots, n-1$

$$\text{return } \beta = \begin{pmatrix} \beta_0(1) & \cdots & \beta_0(n) \\ \beta_1(1) & \cdots & \beta_1(n) \end{pmatrix}$$

---

get_$\gamma(\alpha, \beta)$

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{v=0}^{1} \alpha_v(t)\beta_v(t)} \qquad \text{for } i = 0, 1 \text{ and } t = 1, \ldots, n$$

$$\text{return } \gamma = \begin{pmatrix} \gamma_0(1) & \cdots & \gamma_0(n) \\ \gamma_1(1) & \cdots & \gamma_1(n) \end{pmatrix}$$

---

get_$\xi\left(O^{1,n}, \alpha, \beta, \theta\right)$

$$\xi_{ij}(t) = \frac{\alpha_i(t)a_{ij}b_{j,o_{t+1}}\beta_j(t+1)}{\sum_{v=0}^{1}\sum_{w=0}^{1} \alpha_v(t)a_{vw}b_{w,o_{t+1}}\beta_w(t+1)} \quad \text{for } i, j \in \{0, 1\} \text{ and } t = 1, \ldots, n-1$$

$$\text{return } \xi = (\xi(1) \cdots \xi(n-1)) \text{ where } \xi(t) = \begin{pmatrix} \xi_{00}(t) & \xi_{01}(t) \\ \xi_{10}(t) & \xi_{11}(t) \end{pmatrix}$$

---

get_new_estimates$\left(O^{1,n}, \xi, \gamma\right)$

$$\hat{\pi}_i = \gamma_i(1) \qquad \text{for } i = 0, 1$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{n-1} \xi_{ij}(t)}{\sum_{t=1}^{n-1} \gamma_i(t)} \qquad \text{for } i, j = 0, 1$$

$$\hat{b}_{ik} = \frac{\sum_{t=1}^{n} \delta_k(o_t)\gamma_i(t)}{\sum_{t=1}^{n} \gamma_i(t)} \qquad \text{for } i, k = 0, 1$$

$$\text{where } \delta_k(v) = \begin{cases} 1 & \text{if } v = k \\ 0 & \text{if } v \neq k \end{cases}$$

$$\text{return } \hat{\theta} = (\hat{\pi}, \hat{A}, \hat{B})$$

Note that because of the forward and backward procedures that are executed for each kth iteration for a new set of estimates $\theta^{(k)}$, literally the entire computation has to be done for any additional emission state.

2. Rate of Convergence

Let ϕ be such that $\phi\left(\theta^{(k)}\right) = \theta^{(k+1)}$ where $\theta^{(k)}$ is the kth estimate of θ using Baum-Welch Algorithm. The operator ϕ involves the forward and backward procedures, in which matrices α and β are obtained through iterative computations that go in the different directions of time t. So, ϕ is a non-linear map that cannot be expressed as a combination of ordinary matrix operations, which makes it very difficult, if not impossible, to find its rate of convergence algebraically. Hence, we choose to approximate it experimentally, by observing the linearized behavior of ϕ; i.e., by finding its Jacobian experimentally.

2.1. Jacobian. Utilizing the fact that ϕ is composed of several iterative computations, the Jacobian is computed also iteratively as the B-W Algorithm goes through its own iterations. Here we consider the case when both the transition and observation sequences have a state space $\{0, 1\}$; i.e., the case $m_A = m_B = 2$.

For a notational convenience, let

$$\theta = (r, a, b, x, y)$$

and let

$$\phi(\theta) = (\phi_r(\theta), \phi_a(\theta), \phi_b(\theta), \phi_x(\theta), \phi_y(\theta)),$$

where r, a, b, x, and y, are as defined in equations (1.5), (1.3) and (1.4). What we try to find is the Jacobian

$$\frac{\partial \phi}{\partial \theta} = \begin{vmatrix} \frac{\partial \phi_r}{\partial r} & \frac{\partial \phi_r}{\partial a} & \frac{\partial \phi_r}{\partial b} & \frac{\partial \phi_r}{\partial x} & \frac{\partial \phi_r}{\partial y} \\ \frac{\partial \phi_a}{\partial r} & \frac{\partial \phi_a}{\partial a} & \frac{\partial \phi_a}{\partial b} & \frac{\partial \phi_a}{\partial x} & \frac{\partial \phi_a}{\partial y} \\ \frac{\partial \phi_b}{\partial r} & \frac{\partial \phi_b}{\partial a} & \frac{\partial \phi_b}{\partial b} & \frac{\partial \phi_b}{\partial x} & \frac{\partial \phi_b}{\partial y} \\ \frac{\partial \phi_x}{\partial r} & \frac{\partial \phi_x}{\partial a} & \frac{\partial \phi_x}{\partial b} & \frac{\partial \phi_x}{\partial x} & \frac{\partial \phi_x}{\partial y} \\ \frac{\partial \phi_y}{\partial r} & \frac{\partial \phi_y}{\partial a} & \frac{\partial \phi_y}{\partial b} & \frac{\partial \phi_y}{\partial x} & \frac{\partial \phi_y}{\partial y} \end{vmatrix}.$$

First we rewrite the B-W Algorithm so that finding the partial derivatives can be found easier. For example, as for the forward procedure, which is

$$\begin{aligned} &\alpha_i(1) = \pi_i b_{i,o_1} && \text{for } i = 0, 1, \\ &\alpha_i(t+1) = \left(\alpha_0(t) a_{0i} + \alpha_1(t) a_{1i}\right) b_{i,o_{t+1}} && \text{for } i = 0, 1 \text{ and } t = 2, \ldots, n-1, \end{aligned}$$

we rewrite it as

$$\alpha_{i,1} = \begin{cases} rx & \text{if } i = 0 \text{ and } o_1 = 0 \\ r(1-x) & \text{if } i = 0 \text{ and } o_1 = 1 \\ (1-r)(1-y) & \text{if } i = 1 \text{ and } o_1 = 0 \\ (1-r)y & \text{if } i = 1 \text{ and } o_1 = 1 \end{cases}$$

and, for $t = 2, \ldots, n-1$,

$$\alpha_{i,t+1} = \begin{cases} \left[\, a\, \alpha_{0,t} + (1-b)\, \alpha_{1,t} \,\right] x & \text{if } i = 0 \text{ and } o_{t+1} = 0 \\ \left[\, a\, \alpha_{0,t} + (1-b)\, \alpha_{1,t} \,\right] (1-x) & \text{if } i = 0 \text{ and } o_{t+1} = 1 \\ \left[\, (1-a)\, \alpha_{0,t} + b\, \alpha_{1,t} \,\right] (1-y) & \text{if } i = 1 \text{ and } o_{t+1} = 0 \\ \left[\, (1-a)\, \alpha_{0,t} + b\, \alpha_{1,t} \,\right] y & \text{if } i = 1 \text{ and } o_{t+1} = 1. \end{cases}$$

Note, we also rewrote $\alpha_i(t)$ as $\alpha_{i,t}$ to simplify the notation, since it is rather a function of θ than t now, and note that α is a matrix of functions, each of which elements is a function of θ. Then, we first find $\frac{\partial}{\partial u}\,\alpha$ and $\frac{\partial}{\partial u}\,\beta$, for $u = r, a, b, x, y$, as shown below.

As for $\frac{\partial}{\partial u}\,\alpha$ with $t = 1$, we have

$$\frac{\partial}{\partial r}\,\alpha_{i,1} = \begin{cases} b_{0,o_1} & \text{if } i = 0 \\ -b_{1,o_1} & \text{if } i = 1 \end{cases}$$

$$\frac{\partial}{\partial a}\,\alpha_{i,1} = 0 \qquad \text{for } i = 0, 1$$

$$\frac{\partial}{\partial b}\,\alpha_{i,1} = 0 \qquad \text{for } i = 0, 1$$

$$\frac{\partial}{\partial x}\,\alpha_{i,1} = \begin{cases} r & \text{if } i = 0 \text{ and } o_1 = 0 \\ -r & \text{if } i = 0 \text{ and } o_1 = 1 \\ 0 & \text{if } i = 1 \end{cases}$$

$$\frac{\partial}{\partial y}\,\alpha_{i,1} = \begin{cases} 0 & \text{if } i = 0 \\ r - 1 & \text{if } i = 1 \text{ and } o_1 = 0 \\ 1 - r & \text{if } i = 1 \text{ and } o_1 = 1 \end{cases}$$

and with $t = 1, \ldots, n-1$, we have

$$\frac{\partial}{\partial r}\,\alpha_{i,t+1} = \left(a_{0i}\frac{\partial}{\partial r}\,\alpha_{0,t} + a_{1i}\frac{\partial}{\partial r}\,\alpha_{1,t}\right) b_{i,o_{t+1}} \qquad \text{if } i = 0, 1$$

$$\frac{\partial}{\partial a}\,\alpha_{i,t+1} = \begin{cases} \left(a_{00}\dfrac{\partial}{\partial a}\,\alpha_{0,t} + \alpha_{0,t} + a_{10}\dfrac{\partial}{\partial a}\,\alpha_{1,t}\right) b_{0,o_{t+1}} & \text{if } i = 0 \\ \left(a_{01}\dfrac{\partial}{\partial a}\,\alpha_{0.t} - \alpha_{0,t} + a_{11}\dfrac{\partial}{\partial a}\,\alpha_{1.t}\right) b_{1,o_{t+1}} & \text{if } i = 1 \end{cases}$$

$$\frac{\partial}{\partial b}\,\alpha_{i,t+1} = \begin{cases} \left(a_{00}\dfrac{\partial}{\partial b}\,\alpha_{0,t} + a_{10}\dfrac{\partial}{\partial b}\,\alpha_{1,t} - \alpha_{1,t}\right) b_{0,o_{t+1}} & \text{if } i = 0 \\ \left(a_{01}\dfrac{\partial}{\partial b}\,\alpha_{0,t} + a_{11}\dfrac{\partial}{\partial b}\,\alpha_{1,t} + \alpha_{1,t}\right) b_{1,o_{t+1}} & \text{if } i = 1 \end{cases}$$

$$\frac{\partial}{\partial x}\,\alpha_{i,t+1} = \begin{cases} a_{00}\,\alpha_{0,t} + a_{10}\,\alpha_{1,t} + f_x(t,0)\,b_{00} & \text{if } i = 0,\ o_{t+1} = 0 \\ -a_{00}\,\alpha_{0,t} - a_{10}\,\alpha_{1,t} + f_x(t,0)\,b_{01} & \text{if } i = 0,\ o_{t+1} = 1 \\ f_x(t,1)\,b_{1,o_{t+1}} & \text{if } i = 1 \end{cases}$$

$$\frac{\partial}{\partial y}\,\alpha_{i,t+1} = \begin{cases} f_y(t,0)\,b_{0,o_{t+1}} & \text{if } i = 0 \\ -a_{01}\,\alpha_{0,t} - a_{11}\,\alpha_{1,t} + f_y(t,1)\,b_{10} & \text{if } i = 1,\ o_{t+1} = 0 \\ a_{01}\,\alpha_{0,t} + a_{11}\,\alpha_{1,t} + f_y(t,1)\,b_{11} & \text{if } i = 1,\ o_{t+1} = 1 \end{cases}$$

$$\text{where } f_u(t,i) = a_{0i}\frac{\partial}{\partial u}\alpha_{0,t} + a_{1i}\frac{\partial}{\partial u}\alpha_{1,t} \quad \text{for } u = x, y. \tag{2.10}$$

Note that, since o_t is known and since what we are computing is $\frac{\partial}{\partial\theta}\,\phi(\theta^{(k)})$ for a fixed $\theta^{(k)}$, the parameter estimates $a = a^{(k)}$, $b = b^{(k)}$ and $b_{i,o_t} = b^{(k)}_{i,o_t}$ are constants. So are the values $\frac{\partial}{\partial u}\,\alpha_i(t) = \frac{\partial}{\partial u}\,\alpha_{i,t} = \frac{\partial}{\partial u}\,\alpha_{i,t}$ for each $t = 1, \ldots, n$.

As for the backward procedure, from which we obtain $\frac{\partial}{\partial u}\,\beta$, if $t = n$, we have

$$\frac{\partial}{\partial u}\beta_{i,n} = 0 \quad \text{for } i = 0, 1 \text{ and } u = r, a, b, x, y$$

because $\beta_i(n) = 1$ for $i = 0, 1$. With $t = 1, \ldots, n-1$, we have

$$\beta_i(t) = a_{i0}b_{0,o_{t+1}}\beta_0(t+1) + a_{i1}b_{1,o_{t+1}}\beta_1(t+1).$$

We first define ψ, for $1 \le t \le n-1$, as

$$\psi_{i,t}(j) = a_{ij} b_{j,o_{t+1}} \beta_j(t+1) = \begin{cases} ax\,\beta_0(t+1) & \text{if } i=0, j=0, o_{t+1}=0 \\ a(1-x)\beta_0(t+1) & \text{if } i=0, j=0, o_{t+1}=1 \\ (1-a)(1-y)\,\beta_1(t+1) & \text{if } i=0, j=1, o_{t+1}=0 \\ (1-a)y\,\beta_1(t+1) & \text{if } i=0, j=1, o_{t+1}=1 \\ (1-b)x\,\beta_0(t+1) & \text{if } i=1, j=0, o_{t+1}=0 \\ (1-b)(1-x)\,\beta_0(t+1) & \text{if } i=1, j=0, o_{t+1}=1 \\ b(1-y)\,\beta_1(t+1) & \text{if } i=1, j=1, o_{t+1}=0 \\ by\,\beta_1(t+1) & \text{if } i=1, j=1, o_{t+1}=1, \end{cases}$$

so that

$$\beta_i(t) = \psi_{i,t}(0) + \psi_{i,t}(1) = \sum_{j=0}^{1} \psi_{i,t}(j).$$

Then, for $t = 1, \ldots, n-1$,

$$\frac{\partial}{\partial u}\,\beta_i(t) = \frac{\partial}{\partial u}\,\beta_{i,t} = \sum_{j=0}^{1} \frac{\partial}{\partial u}\,\psi_{i,t}(j) \quad \text{for } i = 0,1 \text{ and } u = r, a, b, x, y, \tag{2.11}$$

where the values of the partial derivatives $\dfrac{\partial}{\partial u}\,\psi_{i,t}(j)$'s are computed as shown below.

$$\frac{\partial}{\partial r}\,\psi_{i,t}(j) = 0 \quad \text{for } i, j \in \{0,1\}$$

$$\frac{\partial}{\partial a}\,\psi_{i,t}(j) = \begin{cases} b_{0,o_{t+1}} \left(a_{00} \dfrac{\partial}{\partial a}\,\beta_{0,t+1} + \beta_{0,t+1} \right) & \text{if } i=0, j=0 \\ b_{1,o_{t+1}} \left(a_{01} \dfrac{\partial}{\partial a}\,\beta_{1,t+1} - \beta_{1,t+1} \right) & \text{if } i=0, j=1 \\ a_{1j} b_{j,o_{t+1}} \dfrac{\partial}{\partial a}\,\beta_{j,t+1} & \text{if } i=1 \end{cases}$$

$$\frac{\partial}{\partial b}\,\psi_{i,t}(j) = \begin{cases} a_{0j} b_{j,o_{t+1}} \dfrac{\partial}{\partial b}\,\beta_{j,t+1} & \text{if } i = 0 \\ b_{0,o_{t+1}} \left(a_{10} \dfrac{\partial}{\partial b}\,\beta_{0,t+1} - \beta_{0,t+1} \right) & \text{if } i = 1, j = 0 \\ b_{1,o_{t+1}} \left(a_{11} \dfrac{\partial}{\partial b}\,\beta_{1,t+1} + \beta_{1,t+1} \right) & \text{if } i = 1, j = 1 \end{cases}$$

$$\frac{\partial}{\partial x}\,\psi_{i,t}(j) = \begin{cases} a_{i0} \left(b_{00} \dfrac{\partial}{\partial x}\,\beta_{0,t+1} + \beta_{0,t+1} \right) & \text{if } j = 0, o_{t+1} = 0 \\ a_{i0} \left(b_{01} \dfrac{\partial}{\partial x}\,\beta_{0,t+1} - \beta_{0,t+1} \right) & \text{if } j = 0, o_{t+1} = 1 \\ a_{i1}\, b_{1,o_{t+1}} \dfrac{\partial}{\partial x}\,\beta_{1,t+1} & \text{if } j = 1 \end{cases}$$

$$\frac{\partial}{\partial y}\,\psi_{i,t}(j) = \begin{cases} a_{i0} b_{0,o_{t+1}} \dfrac{\partial}{\partial y}\,\beta_{0,t+1} & \text{if } j = 0 \\ a_{i1} \left(b_{10} \dfrac{\partial}{\partial y}\,\beta_{1,t+1} - \beta_{1,t+1} \right) & \text{if } j = 1, o_{t+1} = 0 \\ a_{i1} \left(b_{11} \dfrac{\partial}{\partial y}\,\beta_{1,t+1} + \beta_{1,t+1} \right) & \text{if } j = 1, o_{t+1} = 1. \end{cases} \tag{2.12}$$

Using the values of $\frac{\partial}{\partial u}\,\alpha_{i,t}$ and $\frac{\partial}{\partial u}\,\beta_{i,t}$ obtained in equations (2.10) and (2.11) above, we find $\frac{\partial}{\partial u}\,\gamma(\theta)$ for $u = r, a, b, x, r$ as

$$\frac{\partial}{\partial u}\,\gamma_{i,t} = \frac{\dfrac{\partial}{\partial u}\,\alpha_i(t)\beta_i(t)}{\sum_{k=0}^{1} \alpha_k(t)\beta_k(t)} - \frac{\alpha_i(t)\beta_i(t) \sum_{k=0}^{1} \dfrac{\partial}{\partial u}\,\alpha_k(t)\beta_k(t)}{\left[\sum_{k=0}^{1} \alpha_k(t)\beta_k(t)\right]^2} \quad \text{for } i = 0, 1$$

where

$$\frac{\partial}{\partial u}\,\alpha_k(t)\beta_k(t) = \alpha_{k,t} \frac{\partial}{\partial u}\,\beta_{k,t} + \beta_{k,t} \frac{\partial}{\partial u}\,\alpha_{k,t}. \tag{2.13}$$

As for $\frac{\partial}{\partial u}\,\xi(\theta)$, using the notation ψ defined above, we first rewrite ξ as

$$\xi_{ij}(t) = \xi_{i,j,t} = \frac{\alpha_i(t) a_{ij} b_{j,o_{t+1}} \beta_j(t+1)}{\sum_{v=0}^{1} \sum_{w=0}^{1} \alpha_v(t) a_{vw} b_{w,o_{t+1}} \beta_w(t+1)} = \frac{\alpha_{i,t}\psi_{i,t}(j)}{\sum_{v=0}^{1} \sum_{w=0}^{1} \alpha_{v,t}\psi_{v,t}(w)}.$$

Then, using the values of $\frac{\partial}{\partial u}\,\alpha_{i,t}$, $\frac{\partial}{\partial u}\,\beta_{i,t}$, and $\frac{\partial}{\partial u}\,\psi_{i,t}(j)$ that are obtained in equations (2.10), (2.11), and (2.12) above, we find

$$\frac{\partial}{\partial u}\,\xi_{i,j,t} = \frac{\dfrac{\partial}{\partial u}\,\alpha_{i,t}\psi_{i,t}(j)}{\sum_{v=0}^{1}\sum_{w=0}^{1}\alpha_{v,t}\psi_{v,t}(w)} - \frac{\alpha_{i,t}\psi_{i,t}(j)\sum_{v=0}^{1}\sum_{w=0}^{1}\dfrac{\partial}{\partial u}\,\alpha_{v,t}\psi_{v,t}(w)}{\left[\sum_{v=0}^{1}\sum_{w=0}^{1}\alpha_{v,t}\psi_{v,t}(w)\right]^2},$$

where

$$\frac{\partial}{\partial u}\,\alpha_{i,t}\psi_{i,t}(j) = \alpha_{i,t}\frac{\partial}{\partial u}\,\psi_{i,t}(j) + \psi_{i,t}(j)\frac{\partial}{\partial u}\,\alpha_{i,t} \tag{2.14}$$

for $i, j \in \{0, 1\}$. Note that the values of $\xi_{i,j,t}$ can be obtained during the computation of matrix β. Finally, using equations (2.13) and (2.14), we find the partial derivatives for the Jacobian, for $u = r, a, b, x, y$, as shown below.

$$\begin{aligned}
\frac{\partial\phi_r}{\partial u} &= \frac{\partial}{\partial u}\,\gamma_{0,1} \\
\frac{\partial\phi_a}{\partial u} &= \frac{\sum_{t=1}^{n-1}\frac{\partial}{\partial u}\,\xi_{0,0,t}}{\sum_{t=1}^{n-1}\gamma_{0,t}} - \frac{\sum_{t=1}^{n-1}\xi_{0,0,t}\cdot\sum_{t=1}^{n-1}\frac{\partial}{\partial u}\,\gamma_{0,t}}{\left(\sum_{t=1}^{n-1}\gamma_{0,t}\right)^2} \\
\frac{\partial\phi_b}{\partial u} &= \frac{\sum_{t=1}^{n-1}\frac{\partial}{\partial u}\,\xi_{1,1,t}}{\sum_{t=1}^{n-1}\gamma_{1,t}} - \frac{\sum_{t=1}^{n-1}\xi_{1,1,t}\cdot\sum_{t=1}^{n-1}\frac{\partial}{\partial u}\,\gamma_{1,t}}{\left(\sum_{t=1}^{n-1}\gamma_{1,t}\right)^2} \\
\frac{\partial\phi_x}{\partial u} &= \frac{\sum_{t=1}^{n}\delta_0(o_t)\frac{\partial}{\partial u}\,\gamma_{0,t}}{\sum_{t=1}^{n}\gamma_{0,t}} - \frac{\sum_{t=1}^{n}\delta_0(o_t)\gamma_{0,t}\cdot\sum_{t=1}^{n}\frac{\partial}{\partial u}\,\gamma_{0,t}}{\left(\sum_{t=1}^{n}\gamma_{0,t}\right)^2} \\
\frac{\partial\phi_y}{\partial u} &= \frac{\sum_{t=1}^{n}\delta_1(o_t)\frac{\partial}{\partial u}\,\gamma_{1,t}}{\sum_{t=1}^{n}\gamma_{1,t}} - \frac{\sum_{t=1}^{n}\delta_1(o_t)\gamma_{1,t}\cdot\sum_{t=1}^{n}\frac{\partial}{\partial u}\,\gamma_{1,t}}{\left(\sum_{t=1}^{n}\gamma_{1,t}\right)^2}.
\end{aligned} \tag{2.15}$$

As a small demonstration that the above works, consider $\alpha_0(3)$ with the observation sequence $o^{1,3} = (o_1, o_2, o_3) = (0, 1, 1)$. We will find $\alpha_0(3)$ first, then take the partial derivative with respect to a.

$$\alpha_0(1) = \pi_0 b_{0,o_1} = r b_{00} = rx$$

$$\alpha_1(1) = \pi_1 b_{1,o_1} = (1-r) b_{10} = (1-r)(1-y)$$

$$\alpha_0(2) = (\alpha_0(1)a_{00} + \alpha_1(1)a_{10})b_{0,o_2} = [\alpha_0(1)a + \alpha_1(1)(1-b)]\, b_{01}$$
$$= [rxa + (1-r)(1-y)(1-b)]\,(1-x)$$
$$\alpha_1(2) = (\alpha_0(1)a_{01} + \alpha_1(1)a_{11})b_{1,o_2} = [\alpha_0(1)(1-a) + \alpha_1(1)b]\, b_{11}$$
$$= [rx(1-a) + (1-r)(1-y)b]\, y$$

$$\alpha_0(3) = (\alpha_0(2)a_{00} + \alpha_1(2)a_{10})b_{0,o_3}$$
$$= (\alpha_0(2)a + \alpha_1(2)(1-b))b_{01}$$
$$= \big\{\, [rxa + (1-r)(1-y)(1-b)]\,(1-x)a$$
$$+ [rx(1-a) + (1-r)(1-y)b]\, y(1-b)\big\}\,(1-x)$$

So, the partial derivative of $\alpha_0(3)$ is

$$\frac{\partial}{\partial a}\,\alpha_{0,3} = \frac{\partial}{\partial a}\big\{\, [rxa + (1-r)(1-y)(1-b)]\,(1-x)a$$
$$+ [rx(1-a) + (1-r)(1-y)b]\, y(1-b)\big\}\,(1-x)$$
$$= \Big\{\, [rxa + (1-r)(1-y)(1-b)]\,(1-x) +$$
$$a(1-x)\frac{\partial}{\partial a}\,[rxa + (1-r)(1-y)(1-b)]$$
$$+y(1-b)\frac{\partial}{\partial a}\,[rx(1-a) + (1-r)(1-y)b]\,\Big\}\,(1-x)$$
$$= \big\{\, [rxa + (1-r)(1-y)(1-b)]\,(1-x) + a(1-x)rx + y(1-b)(-rx)\,\big\}\,(1-x).$$

Meanwhile the iterative method gives, using the result of the B-W algorithm $\alpha_i(1)$ and $\alpha_i(2)$ for $i = 0, 1$,

$$\frac{\partial}{\partial a}\alpha_{0,1} = 0$$
$$\frac{\partial}{\partial a}\alpha_{1,1} = 0$$

$$\frac{\partial}{\partial a}\,\alpha_{0,2} = \left[a_{00}\frac{\partial}{\partial a}\,\alpha_{0,1} + \alpha_{0,1} + a_{10}\frac{\partial}{\partial a}\,\alpha_{1,1}\right] b_{0,o_2} = rxb_{01} = rx(1-x)$$

$$\frac{\partial}{\partial a}\,\alpha_{1,2} = \left[a_{01}\frac{\partial}{\partial a}\,\alpha_{0,1} - \alpha_{0,1} + a_{11}\frac{\partial}{\partial a}\,\alpha_{1,1}\right] b_{1,o_2} = -rxb_{11} = -rxy$$

$$\begin{aligned}\frac{\partial}{\partial a}\,\alpha_{0,3} &= \left[a_{00}\frac{\partial}{\partial a}\,\alpha_{0,2} + \alpha_{0,2} + a_{10}\frac{\partial}{\partial a}\,\alpha_{1,2}\right] b_{0,o_3} \\ &= \left\{\,[rxa + (1-r)(1-y)(1-b)]\,(1-x) + arx(1-x) + (1-b)(-rxy)\right\}(1-x).\end{aligned}$$

Obviously, the values obtained are the same.

CHAPTER 3

Least Square Estimation

The least square error (LSE) estimation of the parameter set θ is obtained by computing the expected value of θ given an observation sequence $O^{1,n} = (O_1, O_2, \ldots, O_n)$.

$$\theta_{LS} = E\left(\theta \mid O^{1,n}\right) = \int \theta \, P\left(\theta \mid O^{1,n}\right) d\theta.$$

The expected value of the parameter given an observation sequence can be expressed as

$$\begin{aligned}
E\left(\theta \mid O^{1,n}\right) &= \int \theta \, P\left(\theta \mid O^{1,n}\right) d\theta \\
&= \int \theta \frac{P\left(O^{1,n} \mid \theta\right) \; P(\theta)}{P\left(O^{1,n}\right)} d\theta \\
&= \frac{1}{P\left(O^{1,n}\right)} \int \theta \, P\left(O^{1,n} \mid \theta\right) \; P(\theta) d\theta \\
&= \frac{1}{P\left(O^{1,n}\right)} \int \theta \, P\left(O^{1,n} \mid \theta\right) d\theta \qquad \text{assuming } P(\theta) = 1 \\
&= \frac{1}{P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} \int \theta \, P\left(s^{1,n}, O^{1,n} \mid \theta\right) d\theta
\end{aligned}$$

where Ω_n is the set of all possible values of $S^{1,n}$ as before. Thus, we have

$$\theta_{LS} = E\left(\theta \mid O^{1,n}\right) = \frac{1}{P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} \int \theta \pi_{s_1} b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \qquad (3.1)$$

where

$$P\left(O^{1,n}\right) = \sum_{s \in \Omega_n} \int \pi_{s_1} b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta. \qquad (3.2)$$

1. LSE with $m_A = m$, $m_B = 2$

We first consider a general case such that the size of the state space m_A is m, where $m \geq 2$ is any positive integer, and the size of emission space m_B is 2. The transition and emission state spaces are expressed as $\{0, 1, 2, \ldots, m-1\}$ and $\{0, 1\}$, respectively. So, the transition matrix A is an $m \times m$ matrix, and the emission matrix B is an $m \times 2$ matrix as shown below.

$$A = \begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0,m-1} \\ a_{10} & a_{11} & a_{12} & \cdots & a_{1,m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m-1,0} & a_{m-1,1} & a_{m-1,2} & \cdots & a_{m-1,m-1} \end{pmatrix}$$

$$= \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0,\,m-2} & c_{0,\,m-1} \\ c_{1,\,m-1} & c_{10} & \cdots & c_{1,\,m-3} & c_{1,\,m-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{m-2,\,2} & c_{m-2,\,3} & \cdots & c_{m-2,\,0} & c_{m-2,\,1} \\ c_{m-1,\,1} & c_{m-1,\,2} & \cdots & c_{m-1,\,m-1} & c_{m-1,\,0} \end{pmatrix}$$

where $c_{i,j_1} = a_{i,j_2}$ if and only if $j_2 \equiv i + j_1$ in modulo m, and

$$B = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \\ \vdots & \vdots \\ b_{m-1,\,0} & b_{m-1,\,1} \end{pmatrix}.$$

As for the matrix A, the notation c_{ij} is introduced so that the diagonal elements are in the form c_{i0}, $i = 0, 1, \ldots, m-1$, and the parameters to be estimated will be expressed as c_{ij} and b_{jk} where $j = 0, 1, \ldots, m-2$ and $k = 0, 1$. (Note $c_{i,m-1}$ and b_{i1} need not to be estimated

because they are found by $c_{i,m-1} = 1 - \sum_{j=0}^{m-2} c_{i,j}$ and $b_{i1} = 1 - b_{i0}$.) Using this notation, the estimates will include all the diagonal elements of A, by which we later use to reduce a symmetry in the probability distribution. As for the probability distribution for the initial state S_1, let π be as follows:

$$\pi = (\pi_0, \pi_1, \ldots, \pi_{m-1})$$

Since the sum of the probabilities in the above is 1, once the estimates $\hat{\pi}_0, \hat{\pi}_1, \ldots, \hat{\pi}_{m-2}$ are found, π_{m-1} is just $1 - \sum_{i=0}^{m-2} \pi_i$.

1.1. Derivation. Let the initial probabilities $\pi_0 = \pi_1, = \cdots = \pi_{m-2} = \pi_{m-1} = \frac{1}{m}$. It is easy to change the algorithm so that the LSE would also estimate π_i for each i; but, for the simplicity, we do not do it in this paper. With $\pi_i = \frac{1}{m}$, we get

$$\theta_{LS} = E\left(\theta \mid O^{1,n}\right) = \frac{1}{mP\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} \int \theta b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \tag{3.3}$$

where

$$P\left(O^{1,n}\right) = \frac{1}{m} \sum_{s^{1,n} \in \Omega_n} \int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta. \tag{3.4}$$

1.1.1. *Symmetry.* Now there are $m!$ ways to identify the transition states with the numbers $0, 1, \ldots, m-1$. In order to avoid "averaging up" the probability distribution by equivalent states, we compute the LSE under a condition $a_{ii} \geq a_{i+1,i+1}$, or equivalently $c_{i0} \geq c_{i+1,0}$, for all i in $\{0, 1, \ldots, m-2\}$. (Note $a_{ii} = c_{i0}$ for $i = 0, 1, \ldots, m-1$.) The symmetry issue is easier to be understood with state space size $m_A = 2$; so, it is described in more details in Section 2.

1.1.2. *Evaluating Integrals.* The integrals that appear in equations (3.3) and (3.4) can be evaluated if we find the formula for the integral

$$\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta.$$

Let k_{ij} be the number of times the event $S_{t-1} = i$ and $S_t = j$ occurs in a state sequence $S^{1,n} = (S_1, S_2, \ldots, S_n)$ for $2 \leq t \leq n$, and let l_{ij} be the number of times the event $S_t = i$ and $O_t = j$ occurs given the observation sequence $O^{1,n} = (O_1, O_2, \ldots, O_n)$ for $1 \leq t \leq n$. Also, let $K = \{k_{ij}\}$ and $L = \{l_{iu}\}$, $i, j \in \{0, 1, \ldots, m-1\}$, $u \in \{0, 1\}$, denote the set of those values. Then the integral above can be expressed as

$$\int a_{00}^{k_{00}}\, a_{01}^{k_{01}} \cdots a_{0,\,m-1}^{k_{0,\,m-1}} \cdots a_{m-1,\,m-1}^{k_{m-1,\,m-1}}\, b_{00}^{l_{00}}\, b_{01}^{l_{01}} \cdots b_{11}^{l_{11}}\, d\theta.$$

Furthermore, in order to have a better correspondence in the subscript of c_{ij} and those of the exponents that appears as a power to c_{ij} in these two integrals, we define an alternative notation for k_{ij}, which is $d_{i,j_1} = k_{i,j_2}$ if and only if $j_1 \equiv j_2 - i \bmod m$, which gives

$$\begin{pmatrix} k_{00} & k_{01} & \cdots & k_{0,\,m-1} \\ k_{10} & k_{11} & \cdots & k_{1,\,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ k_{m-1,\,0} & k_{m-1,\,1} & \cdots & k_{m-1,\,m-1} \end{pmatrix} = \begin{pmatrix} d_{00} & d_{01} & \cdots & d_{0,\,m-1} \\ d_{1,\,m-1} & d_{1,0} & \cdots & d_{1,\,m-2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m-1,\,1} & d_{m-1,\,2} & \cdots & d_{m-1,\,0} \end{pmatrix}.$$

Now d_{ij} is the exponent for c_{ij} for all $i, j \in \{0, 1, \cdots, m-1\}$. Then, for $m_A > 2$ the integration is in the form

$$\begin{aligned} &\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n}\, d\theta \\ &\qquad = \int_{\hat{A}} \left[c_{00}^{d_{00}}\, c_{01}^{d_{01}} \cdots\, c_{0,\,m-2}^{d_{0,\,m-2}} \left(1 - \sum_{i=0}^{m-2} c_{0\,i}\right)^{d_{0,\,m-1}} \right] \\ &\qquad\qquad \cdots \left[c_{m-1,\,0}^{d_{m-1,\,0}}\, c_{m-1,\,1}^{d_{m-1,\,1}} \cdots\, c_{m-1,\,m-2}^{d_{m-1,\,m-2}} \left(1 - \sum_{i=0}^{m-2} c_{m-1,\,i}\right)^{d_{m-1,\,m-1}} \right] d\hat{A} \\ &\qquad\qquad\qquad \cdot \prod_{i=0}^{m-1} \int_0^1 b_{i0}^{l_{i0}} (1 - b_{i0})^{l_{i1}}\, db_{i0} \end{aligned} \tag{3.5}$$

where $\hat{A} = \{c_{ij} \mid i = 0, 1, \ldots, m-1, j = 0, 1, \ldots, m-2\}$. The value of these integrals can be found by rewriting it using factors and factorials. Let $L = \{l_{ij}\}$, $i \in \{0, 1, \ldots, m-1\}$,

$j \in \{0, 1\}$. Finding the product of integrations over b_{i0} are simple, and we denote it as f, as shown below,

$$f = f(L) = \prod_{i=0}^{m-1} \int_0^1 b_{i0}^{l_{i0}} (1 - b_{i0})^{l_{i1}} db_{i0} = \prod_{i=0}^{m-1} \frac{l_{i0}!\, l_{i1}!}{(l_{i0} + l_{i1} + 1)!}. \tag{3.6}$$

The last equality is obtained using the identity (3.9). Meanwhile, the integration over $\hat{A}$ is not that simple, especially in the derivation. However, the resulting formula is simple enough to use for any value of m_A.

As for the integration over $\hat{A}$, we integrate row-wise; i.e., we first work on each parts inside the square brackets in equation (3.5), then at the end combine the results to evaluate the whole expression. Below is the list of identities that can be found using the binomial theorem, which we use to find the formula for the integration over $\hat{A}$.

$$\int_0^v u^a (v - u)^b du = \sum_{i=0}^{b} \frac{(-1)^i b!}{i!\,(b - i)!\,(a + i + 1)}\, v^{a+b+1} \tag{3.7}$$

$$\int_0^v u^a (1 - u)^b du = \sum_{i=0}^{b} \frac{(-1)^i b!}{i!\,(b - i)!\,(a + i + 1)}\, v^{a+i+1}, \quad \text{and} \tag{3.8}$$

$$\int_0^1 u^a (1 - u)^b du = \frac{a!\, b!}{(a + b + 1)!}, \tag{3.9}$$

for positive integers a and b. All the identities above can be easily verified using the binomial theorem.

In order to simplify the final expression, for $m_A = m > 3$, we first define the following notations for the exponents and sets of indices:

$$\begin{cases} p_k(-1) = d_{k,\,m-1} \\ p_k(i) = p_k(i-1) + d_{k,\,m-i-2} + 1 \quad \text{for } i = 0, 1, \ldots, m-3 \end{cases} \tag{3.10a}$$

$$\begin{cases} \phi_k(I_k) = \phi_{k+1}(I_{k+1}) + d_{k0} + i_k + 1 \quad \text{for } k = 1, 2, \ldots, m-1 \\ \phi_m(I_m) = 0 \end{cases} \tag{3.10b}$$

and

$$I_k = (i_k, i_{k+1}, \ldots, i_{m-2}, i_{m-1}). \tag{3.10c}$$

(Actually for $m_A = 2, 3$, we just do not need to have $p_k(i)$.) Note $p_k(i)$ and $\phi_k(I_k)$ can also be expressed as

$$p_k(i) = \sum_{j=m-i-2}^{m-1} d_{kj} + i + 1 \qquad \text{and} \qquad \phi_k(I_k) = \sum_{j=k}^{m-1} (d_{j0} + i_j) + m - k. \tag{3.11}$$

In addition, as a special case of $p_k(i)$, define $\tilde{p}_k$ as

$$\tilde{p}_k = p_k(m-3) = \sum_{j=1}^{m-1} d_{kj} + m - 2. \tag{3.12}$$

Using the above new notations, we define three functions shown below, which will appear in the final formula, with $k \in \{1, 2, \ldots, m-1\}$, when the identities (3.7), (3.8), and (3.9) are applied.

$$g_{k0}(i,j) = \frac{(-1)^i p_k(j)!}{i!(p_k(j)-i)!(d_{k,\,m-j-3} + i + 1)} \tag{3.13a}$$

$$g_{k1}(I_k) = \frac{(-1)^{i_k} \tilde{p}_k!}{i_k!(\tilde{p}_k - i_k)!\phi_k(I_k)} \tag{3.13b}$$

$$g_2(I_1) = \frac{(\phi_1(I_1) + d_{00})!\,\tilde{p}_0!}{(\phi_1(I_1) + d_{00} + \tilde{p}_0 + 1)!} \tag{3.13c}$$

Note that, applying the identities, we have

$$\int_0^v u^{\,d_{k,\,m-j-3}} (v-u)^{p_k(j)} du = \sum_{i=0}^{p_i(j)} g_{k0}(i,j)\, v^{p_k(j+1)},$$

$$\int_0^v u^{\,\phi_{k+1}(I_{k+1}) + d_{k0}} (1-u)^{\tilde{p}_k} du = \sum_{i_k=0}^{\tilde{p}_k} g_{k1}(I_k)\, v^{\phi_k(I_k)}, \quad \text{and}$$

$$\int_0^1 u^{\,\phi_1(I_1) + d_{00}} (1-u)^{\tilde{p}_0} du = g_2(I_1).$$

Now, since the integration with respect to the parameters of the same row (c_{k0}, c_{k1}, …, $c_{k,\,m-2}$) are in a similar form for each k, we first define ψ_k for each row k, which corresponds to each of the expressions inside square brackets in the equation (3.5).

Let $\psi_m = 1$ and define $c_{-1,\,0}$ as 1, then define a function ψ_k as below for $k = 0, \ldots, m-1$.

$$\psi_k(c_{k-1,\,0}) = \int_0^{c_{k-1,\,0}} \psi_{k+1}(c_{k0})\; c_{k0}^{d_{k0}} \int_0^{1-c_{k0}} c_{k1}^{d_{k1}} \int_0^{1-c_{k0}-c_{k1}} c_{k2}^{d_{k2}} \cdots \int_0^{1-\sum_{j=0}^{m-4} c_{kj}} c_{k,\,m-3}^{d_{k,\,m-3}} \cdot \int_0^{1-\sum_{j=0}^{m-3} c_{kj}} c_{k,\,m-2}^{d_{k,\,m-2}} \left(1 - \sum_{j=0}^{m-2} c_{kj}\right)^{d_{k,\,m-1}} dc_{k,\,m-2}\, dc_{k,\,m-3} \cdots dc_{k1}\, dc_{k0}. \tag{3.14}$$

Then $\psi_0(c_{-1,\,0}) = \psi_0(1)$ is the integral over $\hat{A}$ in equation (3.5); i.e.,

$$\psi_0(1) = \int_{\hat{A}} \left[c_{00}^{d_{00}}\; c_{01}^{d_{01}} \cdots\; c_{0,\,m-2}^{d_{0,\,m-2}} \left(1 - \sum_{i=0}^{m-2}\right)^{d_{0,\,m-1}} \right] \cdots \left[c_{m-1,\,0}^{d_{m-1,\,0}}\; c_{m-1,\,1}^{d_{m-1,\,1}} \cdots\; c_{m-1,\,m-2}^{d_{m-1,\,m-2}} \left(1 - \sum_{i=0}^{m-2}\right)^{d_{m-1,\,m-1}} \right] d\hat{A}$$

so that

$$\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta = f \cdot \psi_0(1).$$

Using the identity (3.7) and the definition of g_{k0} (3.13a), we first simplify the last integral, the one over $c_{k,\,m-2}$, for all $k = 0, 2, \ldots, m-1$.

$$\int_0^{1-\sum_{j=0}^{m-3} c_{kj}} c_{k,\,m-2}^{d_{k,\,m-2}} \left(1 - \sum_{j=0}^{m-3} c_{kj} - c_{k,\,m-2}\right)^{d_{k,\,m-1}} dc_{k,\,m-2} = \sum_{i=0}^{p_k(-1)} g_{k0}(i,-1) \left(1 - \sum_{j=0}^{m-3} c_{kj}\right)^{p_k(0)}.$$

Then, after moving $\sum_{i=0}^{p_k(-1)} g_{k0}(i,-1)$ in front, $\psi_k(c_{k-1,\,0})$ is now expressed as

$$\psi_k(c_{k-1,\,0}) = \sum_{i=0}^{p_k(-1)} g_{k0}(i,-1) \int_0^{c_{k-1,\,0}} \psi_{k+1}(c_{k0})\; c_{k0}^{d_{k0}} \int_0^{1-c_{k0}} c_{k1}^{d_{k1}} \int_0^{1-c_{k0}-c_{k1}} c_{k2}^{d_{k2}} \cdots \int_0^{1-\sum_{j=0}^{m-4} c_{kj}} c_{k,\,m-3}^{d_{k,\,m-3}} \left(1 - \sum_{j=0}^{m-3} c_{kj}\right)^{p_k(0)} dc_{k,\,m-3}\, dc_{k,\,m-4} \cdots dc_{k1}\, dc_{k0}.$$

Applying the same identity again to the integral with respect to $c_{k,\,m-3}$, we get

$$\int_0^{1-\sum_{j=0}^{m-4} c_{kj}} c_{k,\,m-3}^{d_{k,\,m-3}} \left(1 - \sum_{j=0}^{m-4} c_{kj} - c_{k,\,m-3}\right)^{p_k(0)} dc_{k,\,m-3} = \sum_{i=0}^{p_k(0)} g_{k0}(i,0) \left(1 - \sum_{j=0}^{m-4} c_{kj}\right)^{p_k(1)}.$$

so that $\psi_k(c_{k-1,\,0})$ is now

$$\begin{aligned}\psi_k(c_{k-1,\,0}) = &\sum_{i=0}^{p_k(-1)} g_{k0}(i,-1) \sum_{i=0}^{p_k(0)} g_{k0}(i,0) \int_0^{c_{k-1,\,0}} \psi_{k+1}(c_{k0})\, c_{k0}^{d_{k0}} \int_0^{1-c_{k0}} c_{k1}^{d_{k1}} \int_0^{1-c_{k0}-c_{k1}} c_{k2}^{d_{k2}} \\ &\cdots \int_0^{1-\sum_{j=0}^{m-5} c_{ij}} c_{k,\,m-4}^{d_{k,\,m-4}} \left(1 - \sum_{j=0}^{m-4} c_{kj}\right)^{p_k(1)} dc_{k,\,m-4}\, dc_{k,\,m-5} \cdots dc_{k1}\, dc_{k0} \\ = &\prod_{j=-1}^{0} \sum_{i=0}^{p_k(j)} g_{k0}(i,j) \int_0^{c_{k-1,\,0}} \psi_{k+1}(c_{k0})\, c_{k0}^{d_{k0}} \int_0^{1-c_{k0}} c_{k1}^{d_{k1}} \int_0^{1-c_{k0}-c_{k1}} c_{k2}^{d_{k2}} \\ &\cdots \int_0^{1-\sum_{j=0}^{m-5} c_{ij}} c_{k,\,m-4}^{d_{k,\,m-4}} \left(1 - \sum_{j=0}^{m-4} c_{kj}\right)^{p_k(1)} dc_{k,\,m-4}\, dc_{k,\,m-5} \cdots dc_{k1}\, dc_{k0}.\end{aligned}$$

It is easy to observe that, if we apply the identity (3.7) repeatedly, for additional $m-4$ times to the above, we should have

$$\psi_k(c_{k-1,\,0}) = \prod_{j=-1}^{m-4} \sum_{i=0}^{p_k(j)} g_{k0}(i,j) \cdot \int_0^{c_{k-1,\,0}} \psi_{k+1}(c_{k0})\, c_{k0}^{d_{k0}} (1-c_{k0})^{\tilde{p}_k} dc_{k0}$$

for $k = 0, \ldots, m-1$ at the end, to which the identity (3.7) no longer apply. Now it is time to combine all the results. Starting from ψ_{m-1}, we apply the identity (3.8) to the integral in ψ_k for $k = m-1$ down to $k = 1$. With $k = m-1$, we have

$$\begin{aligned}\psi_{m-1}(c_{m-2,\,0}) &= \prod_{j=-1}^{m-4} \sum_{i=0}^{p_{m-1}(j)} g_{m-1,\,0}(i,j) \cdot \int_0^{c_{m-2,\,0}} c_{m-1,\,0}^{d_{m-1,\,0}} (1-c_{m-1,\,0})^{\tilde{p}_{m-1}} dc_{m-1,\,0} \\ &= \prod_{j=-1}^{m-4} \sum_{i=0}^{p_{m-1}(j)} g_{m-1,\,0}(i,j) \cdot \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} g_{m-1,\,1}(I_{m-1})\, c_{m-2,\,0}^{\phi_{m-1}(I_{m-1})}.\end{aligned}$$

Note we used the definition $\phi_m = 0$ and the notation $I_k = (i_k, i_{k+1}, \ldots, i_{m-1})$ in (3.10). So, I_{m-1} is simply i_{m-1}. Then, since ψ_{m-2} contains ψ_{m-1} in its expression as

$$\begin{aligned}\psi_{m-2}(c_{m-3,\,0}) = &\prod_{j=-1}^{m-4} \sum_{i=0}^{p_{m-2}(j)} g_{m-2,\,0}(i,j) \\ &\cdot \int_0^{c_{m-3,\,0}} \psi_{m-1}(c_{m-2,\,0})\, c_{m-2,\,0}^{d_{m-2,\,0}} (1-c_{m-2,\,0})^{\tilde{p}_{m-2}} dc_{m-2,\,0}\,,\end{aligned}$$

plugging in ψ_{m-1} and then applying the same identity (3.8) to ψ_{m-2}, we get

$$\begin{aligned}\psi_{m-2}(c_{m-3,\,0}) &= \prod_{k=m-2}^{m-1}\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]\sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} g_{m-1,\,1}(I_{m-1}) \\ &\qquad \cdot \int_0^{c_{m-3,\,0}} c_{m-2,\,0}^{\phi_{m-1}(I_{m-1})+d_{m-2,\,0}}(1-c_{m-2,\,0})^{\tilde{p}_{m-2}}\,dc_{m-2,\,0} \\ &= \prod_{k=m-2}^{m-1}\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]\sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} g_{m-1,\,1}(I_{m-1}) \\ &\qquad \cdot \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} g_{m-2,\,1}(I_{m-2})\, c_{m-3,\,0}^{\phi_{m-2}(I_{m-2})}.\end{aligned}$$

Repeating the process, we will eventually get

$$\begin{aligned}\psi_1(c_{00}) = \prod_{k=1}^{m-1}&\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right] \\ &\cdot \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} g_{m-1,\,1}(I_{m-1}) \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} g_{m-2,\,1}(I_{m-2}) \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{11}(I_1)\, c_{00}^{\phi_1(I_1)}\end{aligned}$$

so that

$$\begin{aligned}\psi_0(1) = \prod_{k=0}^{m-1}&\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right] \\ &\cdot \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} g_{m-1,\,1}(I_{m-1}) \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} g_{m-2,\,1}(I_{m-2}) \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{11}(I_1) \\ &\qquad \cdot \int_0^1 c_{00}^{\phi_1(I_1)+d_{00}}(1-c_{00})^{\tilde{p}_0}\,dc_{00}.\end{aligned}$$

Now, applying the identity (3.9) to this last integral, we finally have the formula for $\psi_0(1)$, which is

$$\begin{aligned}\psi_0(1) = \prod_{k=0}^{m-1}&\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right] \\ &\cdot \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} g_{m-1,\,1}(I_{m-1}) \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} g_{m-2,\,1}(I_{m-2}) \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{11}(I_1)\, g_2(I_1).\end{aligned}$$

Thus, using the definition of f in (3.6) and the definitions in (3.13), we express the integral as

$$\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \\ = f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}(I_{m-1})\, g_{m-2,1}(I_{m-2}) \cdots g_{11}(I_1)\, g_2(I_1), \tag{3.15}$$

where G is defined as

$$G = \prod_{k=0}^{m-1} \left[\prod_{j=-1}^{m-4} \sum_{i=0}^{p_k(j)} g_{k0}(i,j) \right]. \tag{3.16}$$

Note G is a function of the exponents $\{d_{ij}\}$.

We are now ready to evaluate the expressions for θ_{LS} and $P\left(O^{1,n}\right)$ in the equations (3.3) and (3.4), respectively. First, the equation (3.4) is now

$$P\left(O^{1,n}\right) = \frac{1}{m} \sum_{s^{1,n} \in \Omega_n} f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}(I_{m-1})\; g_{m-2,1}(I_{m-2}) \\ \cdots g_{11}(I_1)\; g_2(I_1). \tag{3.17}$$

where $f, G, g_{k0}, g_{k1}, g_2$, and I_k are as defined in the equations (3.6), (3.16), (3.13a), (3.13b), (3.13c), and (3.10c), respectively.

In order to evaluate θ_{LS} in (3.3) for each parameter, we have to find the corresponding factor as in the case of $m_A = 2$. As for the estimates $\hat{c}_{st}$ for $s \in \{0, \ldots, m-1\}$ and $t \in \{0, \ldots, m-2\}$. Because if the factors that are generated by having the extra c_{st} inside the integral for the estimate $\hat{c}_{st}$, we have the following observations: In the expression (3.15),

- replace $\sum_{i=0}^{p_k(j)} g_{k0}(i,j)$, $k \in \{0, 1, \ldots, m-1\}$, by

$$\begin{cases} \sum_{i=0}^{p_k(j)} g_{k0}^{(1)}(i,j) & \text{if } s = k \text{ and } t = m-j-3 \\ \sum_{i=0}^{p_k(j)+1} g_{k0}^{(2)}(i,j) & \text{if } s = k \text{ and } t \in \{m-j-2,\, m-j-1,\, \ldots,\, m-2\}, \end{cases} \tag{3.18}$$

where

$$g_{k0}^{(1)}(i,j) = \frac{d_{k,\,m-j-3}+i+1}{d_{k,\,m-j-3}+i+2}\, g_{k0}(i,j) \quad \text{and}$$
$$g_{k0}^{(2)}(i,j) = \frac{(-1)^i (p_k(j)+1)!}{i!(p_k(j)-i+1)!(d_{k,\,m-j-3}+i+1)},$$

Note: $g_{k0}^{(2)}(i,j)$ is defined without using the value of $g_{k0}(i,j)$) because the summation with $i = p_i(j)+1$ causes the denominator of $g_{k0}(i,j)$ to be zero. Same with the definition of $g_{k1}^{(2)}(I_k)$ below.

- replace $\sum_{i_k=0}^{\tilde{p}_k} g_{k1}(I_k)$, $k \in \{1,2,\ldots,m-1\}$, by

$$\begin{cases} \sum_{i_k=0}^{\tilde{p}_k} g_{k1}^{(1)}(I_k) & \text{if } s \in \{k, k+1, \ldots, m-1\} \text{ and } t = 0 \\ \sum_{i_k=0}^{\tilde{p}_k+1} g_{k1}^{(2)}(I_k) & \text{if } s = k \text{ and } t \in \{1,2,\ldots,m-2\}, \end{cases} \tag{3.19}$$

where

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

and

- replace $g_2(I_1)$ by

$$\begin{cases} g_2^{(1)}(I_1) & \text{if } s \in \{0,1,\ldots,m-1\} \text{ and } t = 0 \\ g_2^{(2)}(I_1) & \text{if } s = 0 \text{ and } t \in \{1,2,\ldots,m-2\}, \end{cases} \tag{3.20}$$

where

$$g_2^{(1)}(I_1) = \frac{\phi_1(I_1)+d_{00}+1}{\phi_1(I_1)+d_{00}+\tilde{p}_0+2}\, g_2(I_1) \quad \text{and}$$
$$g_2^{(2)}(I_1) = \frac{\tilde{p}_0+1}{\phi_1(I_1)+d_{00}+\tilde{p}_0+2}\, g_2(I_1).$$

The functions $g_{k1}(I_k)$, and $g_2(I_1)$ are functions of the exponents d_{ij} and indices. Since the input indices are fixed once the function is chosen, in the expressions for estimates below, those indices are omitted for the simplicity. Also, define G_{st} as

$$G_{st} = \left[\prod_{j=-1}^{m-t-4}\sum_{i=0}^{p_s(j)} g_{s0}(i,j)\right]\left[\prod_{j=m-t-3}\sum_{i=0}^{p_s(m-t-3)} g_{s0}^{(1)}(i,m-t-3)\right]\left[\prod_{j=m-t-2}^{m-4}\sum_{i=0}^{p_s(j)+1} g_{s0}^{(2)}(i,j)\right] \cdot \prod_{\substack{k=0\\k\neq s}}^{m-1}\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right] \tag{3.21}$$

for $s \in \{0,1,\ldots,m-1\}$ and $t \in \{1,2,\ldots,m-2\}$.

Furthermore, if we define any product notations with index from a_1 to a_2 with $a_1 > a_2$ as 1, and use the definition of f in equation (3.6), the estimates for $\hat{c}_{st}$, $s = 0,1,\ldots,m-1$ and $t = 1,2,\ldots,m-2$ can be expressed as follows:

- Case $t = 0$:

$$\hat{c}_{00} = \frac{1}{mP\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\, g_{m-2,1} \cdots g_{11} \cdot g_2^{(1)} \tag{3.22a}$$

  and, for $s \in \{1,2,\ldots,m-1\}$,

$$\hat{c}_{s0} = \frac{1}{mP\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\, g_{m-2,1} \cdots g_{s+1,1}\, g_{s1}^{(1)}\, g_{s-1,1}^{(1)} \cdots g_{11}^{(1)} \cdot g_2^{(1)}. \tag{3.22b}$$

- Case $t \in \{1,2,\ldots,m-2\}$:

$$\hat{c}_{0t} = \frac{1}{mP\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0t} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\, g_{m-2,1} \cdots g_{11} \cdot g_2^{(2)} \tag{3.22c}$$

and, for $s \in \{1, 2, \ldots, m-1\}$,

$$\hat{c}_{st} = \frac{1}{mP\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} f \cdot G_{st} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s+1}=0}^{\tilde{p}_{s+1}} \sum_{i_s=0}^{\tilde{p}_s+1} \sum_{i_{s-1}=0}^{\tilde{p}_{s-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s+1,\,1}\; g_{s1}^{(2)}\; g_{s-1,\,1} \cdots g_{11} \cdot g_2. \tag{3.22d}$$

As for the estimates $\hat{b}_{k0}$, we define $P(k)$ as

$$P_k = \frac{l_{k0} + 1}{l_{k0} + l_{k1} + 2} \tag{3.23}$$

for $k = 0, 1, \ldots, m-1$, and replace f by $P_k \cdot f$ in equation (3.15) to get

$$\hat{b}_{k0} = \frac{1}{mP\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_k \cdot f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{11} \cdot g_2, \tag{3.24}$$

where G is as defined in equation (3.16) above.

1.1.3. *Finding Exponents.* In order to rewrite the exponent sets K and L, we first define δ and γ as before; i.e.,

$$\delta_i(j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad \text{and} \qquad \gamma_u(v) = \begin{cases} 1 & \text{if } u = v \\ 0 & \text{if } u \neq v \end{cases}$$

where $i, j \in \{0, 1, \ldots, m-1\}$ and $u, v \in \{0, 1\}$. As for γ, we can define it as

$$\gamma_i(j) = \begin{cases} 1 - j & \text{if } i = 0 \\ j & \text{if } i = 1 \end{cases} \tag{3.25}$$

since $i \in \{0, 1\}$. Also, we define the coefficient matrix of γ as below.

$$\widehat{R} = \begin{pmatrix} \hat{r}_{00} & \hat{r}_{01} \\ \hat{r}_{10} & \hat{r}_{11} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \tag{3.26}$$

As for δ, we try to find $(m-1)$-th degree polynomials that do the work. Let

$$\delta_i(j) = r_{i0} + r_{i1} j + r_{i2} j^2 + \cdots + r_{i,\,m-1} j^{\,m-1}. \tag{3.27}$$

Then, the matrix equation below should hold.

$$\begin{pmatrix} r_{00} & r_{01} & \cdots & r_{0,m-1} \\ r_{10} & r_{11} & \cdots & r_{1,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m-1,0} & r_{m-1,1} & \cdots & r_{m-1,m-1} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & m-1 \\ 0 & 1 & 2^2 & \cdots & (m-1)^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 2^{m-1} & \cdots & (m-1)^{m-1} \end{pmatrix} = I \tag{3.28}$$

or

$$RM = I$$

$$R = M^{-1}$$

where

$$M = \{m_{ij}\}, m_{ij} = \{(j-1)^{i-1}\} \quad \text{for } 2 \le i \le m, 1 \le j \le m \text{ and } m_{1j} = 1 \text{ for } 1 \le j \le m, \tag{3.29}$$

$$R = \{r_{i-1,j-1}\}, \tag{3.30}$$

and I is an $m \times m$ identity matrix, in which the element of i-th column and j-th row is the value $\delta_{i-1}(j-1)$. But, M is one of Vandermonde matrices since it is in the form

$$V = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ v_0 & v_1 & \cdots & v_{m-1} \\ v_0^2 & v_1^2 & \cdots & v_{m-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ v_0^{m-1} & v_1^{m-1} & \cdots & v_{m-1}^{m-1} \end{pmatrix}$$

where $v_i = i$ for all $0 \le i \le m-1$. The determinant of this matrix, the Vandermonde determinant, is known to be

$$det(V) = \prod_{0 \le j < i \le m-1} (v_i - v_j).$$

In our matrix $v_i - v_j$ is always positive if $j < i$. So, M is invertible for any positive integer m; i.e., coefficients r_{ij} exist for any m. Furthermore, we can find certain characteristics about r_{ij}. Express the matrices M and $R = M^{-1}$ above as

$$M = \begin{pmatrix} 1 & I_{m-1} \\ 0 & M' \end{pmatrix} \quad \text{and} \quad R = \left(\begin{array}{c|ccc} r_{00} & r_{01} & r_{02} & \cdots & r_{0,m-1} \\ \hline r_{10} & & & & \\ r_{20} & & & & \\ \vdots & & R_3 & & \\ r_{m-1,0} & & & & \end{array}\right) = \begin{pmatrix} R_0 & R_1 \\ R_2 & R_3 \end{pmatrix}$$

where I_{m-1} is a row vector of 1's of length $m-1$, M' is a $m-1 \times m-1$ matrix, $R_0 = r_{00}$, R_1 is a row vector $(r_{01}\, r_{02}\, \cdots\, r_{0,m-1})$, R_2 is a column vector $(r_{10}\, r_{20}\, \cdots\, r_{m-1,0})^T$, and R_3 is a $m-1 \times m-1$ matrix. Then

$$\begin{aligned} MR &= \begin{pmatrix} R_0 + \sum_{i=1}^{m-1} r_{i0} & R_1 + I_{m-1}R_3 \\ M'R_2 & M'R_3 \end{pmatrix} \\ &= \left(\begin{array}{c|ccc} \sum_{i=0}^{m-1} r_{i0} & \sum_{i=0}^{m-1} r_{i1} & \cdots & \sum_{i=0}^{m-1} r_{i,m-1} \\ \hline M'R_2 & & M'R_3 & \end{array}\right) \\ &= I, \end{aligned}$$

which implies

$$\sum_{i=0}^{m-1} r_{ij} = 0 \quad \text{for } j = 1, 2, \ldots, m-1$$

or, in particular,

$$r_{0j} = -\sum_{i=1}^{m-1} r_{ij} \quad \text{for } j = 1, 2, \ldots, m-1. \tag{3.31}$$

Also,

$$
\begin{aligned}
RM &= \begin{pmatrix} R_0 & R_0 I_{m-1} + R_1 M' \\ R_2 & R_2 I_{m-1} + R_3 M' \end{pmatrix} \\
&= \left(\begin{array}{c|c} r_{00} & R_0 I_{m-1} + R_1 M' \\ \hline r_{10} & \\ r_{20} & \\ \vdots & R_2 I_{m-1} + R_3 M' \\ r_{m-1,0} & \end{array}\right) \\
&= I,
\end{aligned}
$$

which implies

$$r_{00} = 1 \tag{3.32}$$

and

$$r_{i0} = 0 \quad \text{for } i = 1, 2, \ldots, m-1. \tag{3.33}$$

The goal here is to find the exponents $K = \{k_{ij}\}$, which we also expressed with $\{d_{st}\}$, and $L = \{l_{jk}\}$. We start with finding the formula for K. Given a state sequence $S^{1,n} = (s_1, s_2, \ldots, s_n)$, for any $i, j \in \{0, 1, \cdots, m-1\}$, we can write k_{ij} as

$$
\begin{aligned}
k_{ij} &= \sum_{t=1}^{n-1} \delta_i(s_t)\delta_j(s_{t+1}) \\
&= \sum_{t=1}^{n-1} \left(r_{i0} + r_{i1}s_t + r_{i2}s_t^2 + \cdots + s_t^{m-1}\right)\left(r_{i0} + r_{i1}s_{t+1} + r_{i2}s_{t+1}^2 + \cdots + s_{t+1}^{m-1}\right) \\
&= \sum_{u=0}^{m-1}\sum_{v=0}^{m-1} r_{iu}r_{jv}\sum_{t=1}^{n-1} s_t^u s_{t+1}^v.
\end{aligned}
$$

Separating the terms with $u, v \neq 0$, with $u \neq 0$ and $v = 0$, with $u = 0$ and $v \neq 0$, and with $u = v = 0$, we get

$$k_{ij} = \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{iu}r_{jv}\sum_{t=1}^{n-1} s_t^u s_{t+1}^v + \sum_{u=1}^{m-1} r_{iu}r_{j0}\sum_{t=1}^{n-1} s_t^u + \sum_{v=1}^{m-1} r_{i0}r_{jv}\sum_{t=1}^{n-1} s_{t+1}^v + r_{i0}r_{j0}(n-1).$$

It turns out that, using the properties of r_{ij}, which are written in (3.31), (3.32), and (3.33), we can express k_{i0} and k_{0j}, $i, j \in \{0, 1, \ldots, n-1\}$ as a function of k_{ij}, $i, j \in \{1, 2, \ldots, n-1\}$, and other quantities, which will be later defined. This conversion is used to make it easier to keep track of the value change in K and L during the execution of an algorithm, which finds the estimates.

In order to make it easier to see the relationship between the values of k_{ij}, we define the following notations, which will not appear in the final expressions:

$$n_u = \sum_{t=1}^{n} s_t^u \qquad \text{and} \qquad n_{uv} = \sum_{t=1}^{n-1} s_t^u s_{t+1}^v \tag{3.34}$$

for $u, v = 1, 2, \ldots, n-1$. Then, we have

$$\sum_{t=1}^{n-1} s_{t+1}^v = \sum_{t=1}^{n} s_t^v - s_1^v = n_v - s_1^v \quad \text{and}$$
$$\sum_{t=1}^{n-1} s_t^u = \sum_{t=1}^{n} s_t^u - s_n^u = n_u - s_n^v.$$

Now the summation over time t is implicit so that k_{ij} can be re-written in a simpler way as

$$k_{ij} = \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{iu}r_{jv}n_{uv} + \sum_{u=1}^{m-1} r_{iu}r_{j0}(n_u - s_n^u) + \sum_{v=1}^{m-1} r_{i0}r_{jv}(n_v - s_1^v) + r_{i0}r_{j0}(n-1).$$

But, because of the property (3.33), when $i, j \neq 0$, only the first term is non-zero, and so

$$k_{ij} = \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{iu}r_{jv}n_{uv} \qquad \text{for } i, j \neq 0. \tag{3.35}$$

If we express remaining k_{ij} in K using the above, we will reduce the number of exponents to count. To further simplify the expression, let k_u, $u = 1, 2, \ldots, m-1$, be the number of u's in the state sequence $S^{1,n}$. Then, again by the property (3.33), we have

$$k_u = \sum_{t=1}^{n} \delta_u(s_t) = \sum_{t=1}^{n} \sum_{v=0}^{m-1} r_{uv} s_t^v = \sum_{t=1}^{n} \sum_{v=1}^{m-1} r_{uv} s_t^v = \sum_{v=1}^{m-1} r_{uv} \sum_{t=1}^{n} s_t^v$$

$$k_u = \sum_{v=1}^{m-1} r_{uv} n_v. \tag{3.36}$$

Now, using the properties of r_{ij}, which are (3.31), (3.32), and (3.33), and using the equations (3.35) and (3.36), we can finally rewrite k_{i0} and k_{0j} as a function of k_{ij} with $i, j \neq 0$, k_i, s_1, and s_n. First, we have

$$\begin{aligned}
k_{00} &= \sum_{u=1}^{m-1} \sum_{v=1}^{m-1} r_{0u} r_{0v} n_{uv} + \sum_{u=1}^{m-1} r_{0u} r_{00} (n_u - s_n^u) + \sum_{v=1}^{m-1} r_{00} r_{0v} (n_v - s_1^v) + r_{00} r_{00} (n-1) \\
&= \sum_{u=1}^{m-1} \sum_{v=1}^{m-1} \left(- \sum_{i=1}^{m-1} r_{iu} \right) \left(- \sum_{j=1}^{m-1} r_{jv} \right) n_{uv} + \sum_{u=1}^{m-1} \left(- \sum_{i=1}^{m-1} r_{iu} \right) (n_u - s_n^u) \\
&\qquad + \sum_{v=1}^{m-1} \left(- \sum_{i=1}^{m-1} r_{iv} \right) (n_v - s_1^v) + n - 1 \\
&= \sum_{u=1}^{m-1} \sum_{v=1}^{m-1} \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} r_{iu} r_{jv} n_{uv} + \sum_{i=1}^{m-1} \left(- \sum_{u=1}^{m-1} r_{iu} n_u + \sum_{u=1}^{m-1} r_{iu} s_n^u \right. \\
&\qquad \left. - \sum_{v=1}^{m-1} r_{iv} n_v + \sum_{v=1}^{m-1} r_{iv} s_1^v \right) + n - 1 \\
&= \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} k_{ij} + \sum_{i=1}^{m-1} \left(-2 \sum_{u=1}^{m-1} r_{iu} n_u + \sum_{u=1}^{m-1} r_{iu} (s_1^u + s_n^u) \right) + n - 1 \\
&= \sum_{i=1}^{m-1} \left(\sum_{j=1}^{m-1} k_{ij} - 2k_i + \sum_{u=1}^{m-1} r_{iu} (s_1^u + s_n^u) \right) + n - 1
\end{aligned}$$

Next, for $j \neq 0$, we have

$$\begin{aligned}
k_{0j} &= \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{0u}r_{jv}n_{uv} + \sum_{u=1}^{m-1} r_{0u}r_{j0}(n_u - s_n^u) + \sum_{v=1}^{m-1} r_{00}r_{jv}(n_v - s_1^v) + r_{00}r_{j0}(n-1) \\
&= \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{0u}r_{jv}n_{uv} + \sum_{v=1}^{m-1} r_{jv}(n_v - s_1^v) \\
&= \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} \left(-\sum_{i=1}^{m-1} r_{iu}\right) r_{jv}n_{uv} + \sum_{v=1}^{m-1} r_{jv}n_v - \sum_{v=1}^{m-1} r_{jv}s_1^v \\
&= -\sum_{i=1}^{m-1} k_{ij} + k_j - \sum_{v=1}^{m-1} r_{jv}s_1^v,
\end{aligned}$$

and for $i \neq 0$,

$$\begin{aligned}
k_{i0} &= \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{iu}r_{0v}n_{uv} + \sum_{u=1}^{m-1} r_{iu}r_{00}(n_u - s_n^u) + \sum_{v=1}^{m-1} r_{i0}r_{0v}(n_v - s_1^v) + r_{i0}r_{00}(n-1) \\
&= \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{iu}r_{0v}n_{uv} + \sum_{u=1}^{m-1} r_{iu}(n_u - s_n^u) \\
&= \sum_{u=1}^{m-1}\sum_{v=1}^{m-1} r_{iu}\left(-\sum_{j=1}^{m-1} r_{jv}\right) n_{uv} + \sum_{u=1}^{m-1} r_{iu}n_u - \sum_{u=1}^{m-1} r_{iu}s_n^u \\
&= -\sum_{j=1}^{m-1} k_{ij} + k_i - \sum_{u=1}^{m-1} r_{iu}s_n^u.
\end{aligned}$$

Now, as for $L = \{l_{ij}\}$, we have a similar situation. However, since $m_B = 2$ is fixed in this paper, reducing the number of l_{ij} that need to be found is simpler. As the indicator function γ for an observation sequence, we use the matrix $\widehat{R}$ defined in (3.26), which is

$$\widehat{R} = \begin{pmatrix} \hat{r}_{00} & \hat{r}_{01} \\ \hat{r}_{10} & \hat{r}_{11} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

Given a state sequence $S^{1,n} = (s_1, s_2, \ldots, s_n)$ and an observation sequence $O^{1,n} =$

$(o_1, o_2, \ldots, o_n)$, for any $i \in \{0, 1, \ldots, m-1\}$ and $j \in \{0, 1\}$, we can write l_{ij} as

$$\begin{aligned} l_{ij} &= \sum_{t=1}^{n-1} \delta_i(s_t)\gamma_j(o_t) \\ &= \sum_{t=1}^{n-1} \left(r_{i0} + r_{i1}s_t + r_{i2}s_t^2 + \cdots + s_t^{m-1}\right)\left(\hat{r}_{j0} + \hat{r}_{j1}o_t\right) \\ &= \sum_{u=0}^{m-1}\sum_{v=0}^{1} r_{iu}\hat{r}_{jv}\sum_{t=1}^{n} s_t^u o_t^v \end{aligned}$$

Separating the terms with $u \neq 0$ and $v = 1$, with $u \neq 0$ and $v = 0$, with $u = 0$ and $v = 1$, and with $u = v = 0$, we get

$$l_{ij} = \sum_{u=1}^{m-1} r_{iu}\hat{r}_{j1}\sum_{t=1}^{n} s_t^u o_t + \sum_{u=1}^{m-1} r_{iu}\hat{r}_{j0}\sum_{t=1}^{n} s_t^u + r_{i0}\hat{r}_{j1}\sum_{t=1}^{n} o_t + r_{i0}\hat{r}_{j0}n.$$

Again, to simplify the expressions, we introduce a notation m_{u1}, which will not appear in the final expression, as

$$m_{u1} = \sum_{i=1}^{n} s_t^u o_t. \tag{3.37}$$

Also, let l_1 be the number of 1's in the observation sequence $O^{1,n}$. Then, since $o_t \in \{0, 1\}$, we have

$$l_1 = \sum_{t=1}^{n} o_t. \tag{3.38}$$

Using the notations m_{u1} and l_1, l_{ij} is now

$$l_{ij} = \sum_{u=1}^{m-1} r_{iu}\hat{r}_{j1}m_{u1} + \sum_{u=1}^{m-1} r_{iu}\hat{r}_{j0}n_u + r_{i0}\hat{r}_{j1}l_1 + r_{i0}\hat{r}_{j0}n. \tag{3.39}$$

We see that if $i \neq 0$ and $j = 1$, since we know the exact value of $\widehat{R}$, after using the property (3.33), the expression is very simple as shown below.

$$l_{i1} = \sum_{u=1}^{m-1} r_{iu}m_{u1} \qquad \text{for } i \neq 0. \tag{3.40}$$

We are now ready to express l_{00}, l_{01}, and l_{i0} for $i \in \{1, 2, \ldots, m-1\}$ as functions of l_{i1}, k_i, and l_1 as shown below. By the equation (3.40), (3.36), the properties (3.31), (3.32)

$$\begin{aligned}
l_{00} &= \sum_{u=1}^{m-1} r_{0u}\hat{r}_{01}m_{u1} + \sum_{u=1}^{m-1} r_{0u}\hat{r}_{00}n_u + r_{00}\hat{r}_{01}l_1 + r_{00}\hat{r}_{00}n \\
&= -\sum_{u=1}^{m-1}\left(-\sum_{i=1}^{m-1} r_{iu}\right)m_{u1} + \sum_{u=1}^{m-1}\left(-\sum_{i=1}^{m-1} r_{iu}\right)n_u - l_1 + n \\
&= \sum_{i=1}^{m-1}(l_{i1} - k_i) - l_1 + n, \\
l_{01} &= \sum_{u=1}^{m-1} r_{0u}\hat{r}_{11}m_{u1} + \sum_{u=1}^{m-1} r_{0u}\hat{r}_{10}n_u + r_{00}\hat{r}_{11}l_1 + r_{00}\hat{r}_{10}n \\
&= \sum_{u=1}^{m-1}\left(-\sum_{i=1}^{m-1} r_{iu}\right)m_{u1} + l_1 \\
&= -\sum_{i=1}^{m-1} l_{i1} + l_1, \qquad \text{and} \\
l_{i0} &= \sum_{u=1}^{m-1} r_{iu}\hat{r}_{01}m_{u1} + \sum_{u=1}^{m-1} r_{iu}\hat{r}_{00}n_u + r_{i0}\hat{r}_{01}l_1 + r_{i0}\hat{r}_{00}n \\
&= -l_{i1} + k_i.
\end{aligned}$$

In summary, for $i, j \neq 0$, we have

$$k_{00} = \sum_{i=1}^{m-1}\left[\sum_{j=1}^{m-1} k_{ij} - 2k_i + \left(\sum_{j=1}^{m-1} r_{ji}\right)(s_1^i + s_n^i)\right] + n - 1, \tag{3.41a}$$

$$k_{0j} = -\sum_{i=1}^{m-1}\left(k_{ij} + r_{ji}s_1^i\right) + k_j, \tag{3.41b}$$

$$k_{i0} = -\sum_{j=1}^{m-1}\left(k_{ij} + r_{ij}s_n^j\right) + k_i \tag{3.41c}$$

$$l_{00} = \sum_{i=1}^{m-1}(l_{i1} - k_i) - l_1 + n, \tag{3.41d}$$

$$l_{01} = -\sum_{i=1}^{m-1} l_{i1} + l_1, \quad \text{and} \tag{3.41e}$$

$$l_{i0} = -l_{i1} + k_i. \tag{3.41f}$$

Now, all the elements in K and L can be expressed as a function of ω, which is defined as

$$\omega = \left\{K_{m-1}, K^{m-1,\,m-1}, L_{m-1}, s_1, s_n\right\}$$

where

$$K_{m-1} = (k_1,\, k_2,\, \ldots,\, k_{m-1})\,,$$

$$K^{m-1,m-1} = \begin{pmatrix} k_{11} & k_{12} & \cdots & k_{1,\,m-1} \\ k_{21} & k_{22} & \cdots & k_{2,\,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ k_{m-1,\,1} & k_{m-1,\,2} & \cdots & k_{m-1,\,m-1} \end{pmatrix}, \qquad \text{and}$$

$$L_{m-1} = (l_{11},\, l_{21},\, \ldots,\, l_{m-1,\,1})\,. \tag{3.42}$$

Note that n and l_1 is omitted because they are constant, given $O^{1,n}$.

Let $h_n(\omega)$ be the number of sequence(s) $s^{1,\,n} \in \Omega_n$ that generates the same value of ω given a particular $O^{1,\,n}$, where Ω_n is the set of all sequences that are composed of $0, 1, \ldots, m-1$ and of length n. We define $\overline{\Omega}$ as

$$\overline{\Omega}_n(o^{1,n}) = \{\omega \mid \omega \text{ corresponds to some } s^{1,n} \in \Omega_n \text{ and } o^{1,n}\}. \tag{3.43}$$

Not all possible values of ω can happen under the restriction on $S^{1,n}$ and given a particular $O^{1,n}$; so, the size of $\overline{\Omega}_n$ is smaller than m^n, the size of Ω_n. An example is shown in Section 2. Also, we can see from the definition of ω that the size of $\overline{\Omega}_n$ is smaller than $m^2 n^{2(m-1)+(m-1)^2} < cm^2 n^{(m-1)^2}$ for some $c \ll 1$. Experimentally, we can see the size of $\overline{\Omega}_n$ is considerably smaller, which will be shown in Chapter 4.

The algorithm for finding all the elements in $\overline{\Omega}_n$ and the value of its corresponding h_n is based on the following observation:

- If the next transition state s_{t+1} is 0, then

  - the last state s_n will be 0, and
  - the rest of the quantities will stay the same

- If the next transition state s_{t+1} is j, where $j \in \{1, 2, \ldots, m-1\}$, then
  - the last state s_n will be j,
  - k_j (the number of j's in $s^{1,t+1}$) will be increased by 1,
  - k_{ij} (the number of ij's in $s^{1,t+1}$) will be increased by 1 if the current state s_t is i, and will stay the same otherwise, for $i \in \{1, 2, \ldots, m-1\}$,
  - l_{j1} (the number of times 1 is observed when the state is j, given $s^{1,t+1}$ and $o^{1,t+1}$) will be increased by 1 if $o_{t+1} = 1$, and stays the same otherwise, and
  - the rest of the quantities will stay the same.

However, because of the symmetries in the system, the values of h_n are distributed symmetrically for each value of s_1; i.e., once we get the values of h_n that correspond to all sequences $s^{1,n}$ that starts with, say, $s_1 = 0$, then we can find the rest of values of h_n for $s_1 = 1, 2, \ldots, m-1$ by just interchanging the subscripts of k_{ij} and l_{ij} accordingly. For example, if we want to find h_n-values that corresponds to the set of state sequences that starts with $S_1 = 2$, while we already have all the values of h_n-values that corresponds to the set of state sequences that starts with $S_1 = 0$, then let $k_{i2} = x$, $k_{2j} = y$, and $l_{2k} = z$, if $k_{i0} = x$, $k_{0j} = y$, and $l_{0k} = z$, respectively (i.e., interchange the first and third rows and columns of the matrix $\{k_{ij}\}$ and the first and third rows of the matrix $\{l_{ik}\}$). A simpler example is shown in Section 2 with $m_A = 2$.

Thus, the algorithm below only finds h_n that corresponds to the set of state sequences $S^{1,n}$ that starts with $S_1 = 0$.

Algorithm for finding h_n

Let $h_1(\omega_0) = 1$, where ω_0 is an ω-value such that all the entries in the elements in ω is 0.

for t from 1 to $n-1$

with all $\omega = (K_{m-1}, K^{m-1,\,m-1}, L_{m-1}, 0, s_t)$ such that $h_t(\omega) > 0$

Increment $h_{t+1}(K_{m-1}, K^{m-1,\,m-1}, L_{m-1}, 0, 0)$ by the value $h_t(\omega)$ (for the case $s_{t+1} = 0$).

for s_{t+1} from 1 to $m-1$

Obtain $\hat{\omega}$ from ω by incrementing

(i) $k_{s_{t+1}}$ in K_{m-1} by one,

(ii) $k_{s_t,\,s_{t+1}}$ in $K^{m-1\,m-1}$ by one, and

(iii) $l_{s_{t+1},1}$ in L_{m-1} by O_{t+1},

then by letting S_t take the value S_{t+1}.

Increment $h_{t+1}(\hat{\omega})$ by the value $h_t(\omega)$.

end for

end for

Since the algorithm goes through a for-loop $n-1$ times, the value of S_1 is fixed as zero, and the size of $\overline{\Omega}_n$ is less than $cm^2n^{(m-1)^2}$ for some constant c, the computational complexity is less than $d\,mn^{m^2}$ for some constant d, and experiments have shown that $d \ll 1$.

1.2. Covariance Matrix. For the general case in which the state space m_A is any integer m, and the emission state space is 2, we have a $(m+2)$ by $(m+2)$ covariance matrix, given an observation sequence of length n, $O^{1,n}$. The entries of this matrix are:

$$Cov\left(a_{i_1,j_1}, a_{i_2,j_2} \mid O^{1,n}\right) \qquad \text{for } i_1, j_1, i_2, j_2 \in \{0, 1, \ldots, m-1\}$$

$$Cov\left(b_{i_1,k_1}, b_{i_2,k_2} \mid O^{1,n}\right) \qquad \text{for } i_1, i_2 \in \{0, 1, \ldots, m-1\} \text{ and } k_1, k_2 \in \{0, 1\}, \text{ and}$$

$$Cov\left(a_{i_1,j_1}, b_{i_2,k}) \mid O^{1,n}\right) \qquad \text{for } i_1, j_1, i_2 \in \{0, 1, \ldots, m-1\} \text{ and } k \in \{0, 1\}.$$

Or, using the notation $c_{i,j}$ for $a_{\hat{i},\hat{j}}$, where $c_{i,j_1} = a_{i,j_2}$ if and only if $j_2 \equiv i + j_1$ in modulo m, what we want are

$$Cov\left(c_{i_1,j_1}, c_{i_2,j_2} \mid O^{1,n}\right) \qquad \text{for } i_1, j_1, i_2, j_2 \in \{0, 1, \ldots, m-1\}$$

$$Cov\left(b_{i_1,k_1}, b_{i_2,k_2} \mid O^{1,n}\right) \qquad \text{for } i_1, i_2 \in \{0, 1, \ldots, m-1\} \text{ and } k_1, k_2 \in \{0, 1\}, \text{ and}$$

$$Cov\left(c_{i_1,j_1}, b_{i_2,k}) \mid O^{1,n}\right) \qquad \text{for } i_1, j_1, i_2 \in \{0, 1, \ldots, m-1\} \text{ and } k \in \{0, 1\}.$$

Since

$$Cov\left(u, v \mid O^{1,n}\right) = E\left(uv \mid O^{1,n}\right) - E\left(u \mid O^{1,n}\right) E\left(v \mid O^{1,n}\right), \tag{3.44}$$

and the formulas for $E\left(u \mid O^{1,n}\right)$ and $E\left(v \mid O^{1,n}\right)$ are already found, here we formulate $E\left(uv \mid O^{1,n}\right)$ only. Consider $Cov\left(c_{s_1,s_2} \mid O^{1,n}\right)$, which is in the form

$$\begin{aligned} & E\left(c_{s_1t_1}\, c_{s_2t_2} \mid O^{1,n}\right) \\ & \quad = \int_{\hat{A}} c_{s_1t_1} c_{s_2t_2} \left[c_{00}^{d_{00}}\ c_{01}^{d_{01}} \cdots\ c_{0,\,m-2}^{d_{0,\,m-2}} \left(1 - \sum_{i=0}^{m-2}\right)^{d_{0,\,m-1}} \right] \qquad (3.45) \\ & \qquad \cdots \left[c_{m-1,\,0}^{d_{m-1,\,0}}\ c_{m-1,\,1}^{d_{m-1,\,1}} \cdots\ c_{m-1,\,m-2}^{d_{m-1,\,m-2}} \left(1 - \sum_{i=0}^{m-2}\right)^{d_{m-1,\,m-1}} \right] d\hat{A} \\ & \qquad\qquad \cdot \prod_{i=0}^{m-1} \int_0^1 b_{i0}^{l_{i0}} (1 - b_{i0})^{l_{i1}}\, db_{i0} \qquad (3.46) \end{aligned}$$

where $\hat{A} = \{c_{ij} \mid i = 0, 1, \ldots, m-1, j = 0, 1, \ldots, m-2\}$, as in Equation (3.5). First we define additional functions as follows:

Let

$$g_{k0}^{(11)}(i,j) = \frac{d_{k,\,m-j-3} + i + 1}{d_{k,\,m-j-3} + i + 3}\, g_{k0}(i,j), \tag{3.47a}$$

$$g_{k0}^{(22)}(i,j) = \frac{(-1)^i (p_k(j)+2)!}{i!(p_k(j)-i+2)!(d_{k,\,m-j-3}+i+1)}, \tag{3.47b}$$

$$g_{k1}^{(11)}(I_k) = \frac{\phi_k(I_k)}{\phi_k(I_k)+2}\, g_{k1}(I_k), \tag{3.47c}$$

$$g_{k1}^{(22)}(I_k) = \frac{(-1)^{i_k} (\tilde{p}_k+2)!}{i_k!(\tilde{p}_k - i_k + 2)!\phi_k(I_k)}, \tag{3.47d}$$

$$g_{k1}^{(12)}(I_k) = \frac{(-1)^{i_k} (\tilde{p}_k+1)!}{i_k!(\tilde{p}_k - i_k + 1)!\,(\phi_k(I_k)+1)}, \tag{3.47e}$$

$$g_{2}^{(11)}(I_1) = \frac{\phi_1(I_1) + d_{00} + 2}{\phi_1(I_1) + d_{00} + \tilde{p}_0 + 3}\, g_{2}^{(1)}(I_1), \tag{3.47f}$$

$$g_{2}^{(22)}(I_1) = \frac{\tilde{p}_0 + 2}{\phi_1(I_1) + d_{00} + \tilde{p}_0 + 3}\, g_{2}^{(2)}(I_1), \quad \text{and} \tag{3.47g}$$

$$g_{2}^{(12)}(I_1) = \frac{\tilde{p}_0 + 1}{\phi_1(I_1) + d_{00} + \tilde{p}_0 + 3}\, g_{2}^{(1)}(I_1). \tag{3.47h}$$

where $\tilde{p}_k$ is as defined in equation (3.12). Also, define $G_{s_1,\,t_1,\,s_2,\,t_2}$ as follows:

(1) If $s_1 = s_2 = s$ and $t_1 = t_2 = t$, then

$$G_{s\,t,\,s\,t} = \left[\prod_{j=-1}^{m-t-4} \sum_{i=0}^{p_s(j)} g_{s0}(i,j)\right] \left[\sum_{i=0}^{p_s(m-t-3)} g_{s0}^{(11)}(i, m-t-3)\right] \left[\prod_{j=m-t-2}^{m-4} \sum_{i=0}^{p_s(j)+1} g_{s0}^{(22)}(i,j)\right] \cdot \prod_{\substack{k=0 \\ k\neq s}}^{m-1} \left[\prod_{j=-1}^{m-4} \sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]. \tag{3.48a}$$

(2) If $s_1 = s_2 = s$ and $t_1 \neq t_2$, let $T = max\{t_1, t_2\}$ and $t = min\{t_1, t_2\}$. Then

$$G_{s\,t_1,\,s\,t_2} = \left[\prod_{j=-1}^{m-T-4}\sum_{i=0}^{p_s(j)} g_{s0}(i,j)\right]\left[\sum_{i=0}^{p_s(m-T-3)} g_{s0}^{(1)}(i, m-T-3)\right]\left[\prod_{j=m-T-2}^{m-t-4}\sum_{i=0}^{p_s(j)+1} g_{s0}^{(2)}(i,j)\right]$$
$$\cdot\left[\sum_{i=0}^{p_s(m-t-3)} g_{s0}^{(12)}(i, m-t-3)\right]\left[\prod_{j=m-t-2}^{m-4}\sum_{i=0}^{p_s(j)+1} g_{s0}^{(22)}(i,j)\right]$$
$$\cdot\prod_{\substack{k=0\\k\neq s}}^{m-1}\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]. \tag{3.48b}$$

(3) If $s_1 \neq s_2$ then

$$G_{s_1\,t_1,\,s_2\,t_2} = \left[\prod_{j=-1}^{m-t_1-4}\sum_{i=0}^{p_{s_1}(j)} g_{s_1,\,0}(i,j)\right]\left[\sum_{i=0}^{p_{s_1}(m-t_1-3)} g_{s_1,\,0}^{(1)}(i, m-t_1-3)\right]$$
$$\cdot\left[\prod_{j=m-t_1-2}^{m-4}\sum_{i=0}^{p_{s_1}(j)+1} g_{s_1,\,0}^{(2)}(i,j)\right]\left[\prod_{j=-1}^{m-t_2-4}\sum_{i=0}^{p_{s_2}(j)} g_{s_2,\,0}(i,j)\right]$$
$$\cdot\left[\sum_{i=0}^{p_{s_2}(m-t_2-3)} g_{s_2,\,0}^{(1)}(i, m-t_2-3)\right]\left[\prod_{j=m-t_2-2}^{m-4}\sum_{i=0}^{p_{s_2}(j)+1} g_{s_2,\,0}^{(2)}(i,j)\right]$$
$$\cdot\prod_{\substack{k=0\\k\neq s_1,\,s_2}}^{m-1}\left[\prod_{j=-1}^{m-4}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]. \tag{3.48c}$$

Also, let

$$F_m = \frac{1}{mP\left(O^{1,n}\right)}.$$

Then, extending the formula for the expected values $E\left(c_{s_1,t_1}c_{s_2,t_2} \mid O^{1,n}\right)$, for $s_1, s_2 \in \{1, 2, \ldots, m-1\}$ and $t_1, t_2 \in \{1, 2, \ldots, m-2\}$, we have the following:

$$E\left(c_{00}\,c_{00} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f\cdot G\sum_{i_{m-1}=0}^{\tilde{p}_{m-1}}\sum_{i_{m-2}=0}^{\tilde{p}_{m-2}}\cdots\sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\,g_{m-2,\,1}\cdots g_{11}\cdot g_2^{(11)}. \tag{3.49a}$$

$$E\left(c_{00}\,c_{0t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f\cdot G_{0\,t_2}\sum_{i_{m-1}=0}^{\tilde{p}_{m-1}}\sum_{i_{m-2}=0}^{\tilde{p}_{m-2}}\cdots\sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\,g_{m-2,\,1}\cdots g_{11}\cdot g_2^{(12)}. \tag{3.49b}$$

$$E\left(c_{00}\, c_{s_2 0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s_2+1,\,1}\; g^{(1)}_{s_2,\,1}\; g^{(1)}_{s_2-1,\,1} \cdots g^{(1)}_{11} \cdot g^{(11)}_{2}. \tag{3.49c}$$

$$E\left(c_{00}\, c_{s_2 t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_2\, t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s_2+1}=0}^{\tilde{p}_{s_2+1}} \sum_{i_{s_2}=0}^{\tilde{p}_{s_2}+1} \sum_{i_{s_2-1}=0}^{\tilde{p}_{s_2-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s_2+1,\,1}\; g^{(2)}_{s_2,\,1}\; g_{s_2-1,\,1} \cdots g_{11} \cdot g^{(1)}_{2}. \tag{3.49d}$$

$$E\left(c_{0t_1}\, c_{0t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0\, t_1,\, 0\, t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{11} \cdot g^{(22)}_{2}. \tag{3.49e}$$

$$E\left(c_{0t_1}\, c_{s_2 t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0\, t_1,\, s_2\, t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s_2+1}=0}^{\tilde{p}_{s_2+1}} \sum_{i_{s_2}=0}^{\tilde{p}_{s_2}+1} \sum_{i_{s_2-1}=0}^{\tilde{p}_{s_2-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s_2+1,\,1}\; g^{(2)}_{s_2,\,1}\; g_{s_2-1,\,1} \cdots g_{11} \cdot g^{(2)}_{2}. \tag{3.49f}$$

$$E\left(c_{s_1 0}\, c_{0t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0\, t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s_1+1,\,1}\; g^{(1)}_{s_1,\,1}\; g^{(1)}_{s_1-1,\,1} \cdots g^{(1)}_{11} \cdot g^{(12)}_{2}. \tag{3.49g}$$

Let $Q = max\{s_1, s_2\}$ and $q = min\{s_1, s_2\}$ then

$$E\left(c_{s_1 0}\, c_{s_2 0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{Q+1,\,1}\; g^{(1)}_{Q1}\; g^{(1)}_{Q-1,\,1} \cdots g^{(1)}_{q+1,\,1}\; g^{(11)}_{q1}\; g^{(11)}_{q-1,\,1} \cdots g^{(11)}_{11} \cdot g^{(11)}_{2}. \tag{3.49h}$$

If $s_1 = s_2 = s$ then

$$E\left(c_{s0}\, c_{st_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s\,t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s+1}=0}^{\tilde{p}_{s+1}} \sum_{i_s=0}^{\tilde{p}_s+1} \sum_{i_{s-1}=0}^{\tilde{p}_{s-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s+1,\,1}\; g_{s1}^{(12)}\; g_{s-1,\,1}^{(1)} \cdots g_{11}^{(1)} \cdot g_2^{(1)}. \tag{3.49i}$$

If $s_1 > s_2$ then

$$E\left(c_{s_1 0}\, c_{s_2 t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_2\,t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s_2+1}=0}^{\tilde{p}_{s_2+1}} \sum_{i_{s_2}=0}^{\tilde{p}_{s_2}+1} \sum_{i_{s_2-1}=0}^{\tilde{p}_{s_2-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s_1+1,\,1}\; g_{s_1,\,1}^{(1)}\; g_{s_1-1,\,1}^{(1)} \cdots g_{s_2+1,\,1}^{(1)}\; g_{s_2,\,1}^{(12)}\; g_{s_2-1,\,1}^{(1)} \cdots g_{11}^{(1)} \cdot g_2^{(1)}. \tag{3.49j}$$

If $s_1 < s_2$ then

$$E\left(c_{s_1 0}\, c_{s_2 t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_2\,t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s_2+1}=0}^{\tilde{p}_{s_2+1}} \sum_{i_{s_2}=0}^{\tilde{p}_{s_2}+1} \sum_{i_{s_2-1}=0}^{\tilde{p}_{s_2-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1}\; \cdots g_{s_2+1,\,1}\; g_{s_2,\,1}^{(2)}\; g_{s_2-1,\,1} \cdots g_{s_1+1,\,1}\; g_{s_1,\,1}^{(1)}\; g_{s_1-1,\,1}^{(1)} \cdots g_{11}^{(1)} \cdot g_2^{(1)}. \tag{3.49k}$$

If $s_1 = s_2 = s$ then

$$E\left(c_{s,\,t_1} c_{s,\,t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s,\,t_1,\,s,\,t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s+1}=0}^{\tilde{p}_{s+1}} \sum_{i_s=0}^{\tilde{p}_s+2} \sum_{i_{s-1}=0}^{\tilde{p}_{s-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,\,1}\; g_{m-2,\,1} \cdots g_{s+1,\,1}\; g_{s1}^{(22)}\; g_{s-1,\,1} \cdots g_{11} \cdot g_2. \tag{3.49l}$$

If $s_1 \neq s_2$ then, for $Q = max\{s_1, s_2\}$ and $q = min\{s_1, s_2\}$,

$$E\left(c_{s_1 t_1}\, c_{s_2 t_2} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_1 t_1\, s_2 t_2} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{Q+1}=0}^{\tilde{p}_{Q+1}} \sum_{i_Q=0}^{\tilde{p}_Q(m-3)+1} \sum_{i_{Q-1}=0}^{\tilde{p}_{Q-1}}$$
$$\cdots \sum_{i_{q+1}=0}^{\tilde{p}_{q+1}} \sum_{i_q=0}^{\tilde{p}_q+1} \sum_{i_{q-1}=0}^{\tilde{p}_{q-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\; g_{m-2,1}$$
$$\cdots g_{Q+1,1}\; g_{Q1}^{(2)}\; g_{Q-1,1} \cdots g_{q+1,1}\; g_{q1}^{(2)}\; g_{q-1,1} \cdots g_{11} \cdot g_2.$$

Similarly, as for the expected values $E\left(b_{s_1,0} b_{s_2,0} \mid O^{1,n}\right)$, $s_1, s_2 \in \{0, 1, \ldots, m-1\}$, we define $\widetilde{P}_s$ as

$$\widetilde{P}_s = \frac{l_{s0} + 2}{l_{s0} + l_{s1} + 3}. \tag{3.50}$$

Then, if $s_1 = s_2 = s$, we have

$$E\left(b_{s0}\, b_{s0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} \widetilde{P}_s P_s \cdot f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\; g_{m-2,1} \cdots g_{11} \cdot g_2, \tag{3.51a}$$

and if $s_1 \neq s_2$, then

$$E\left(b_{s_1 0}\, b_{s_2 0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} P_{s_1} P_{s_2} \cdot f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\; g_{m-2,1}$$
$$\cdots g_{11} \cdot g_2. \tag{3.51b}$$

As for the expected values in the form $E\left(c_{s_1 t}\, b_{s_2 0} \mid O^{1,n}\right)$, we have the following:

$$E\left(c_{s_1 0}\, b_{s_2 0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} P_{s_2}\, f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\; g_{m-2,1}$$
$$\cdots g_{s_1+1,1}\; g_{s_1,1}^{(1)}\; g_{s_1-1,1}^{(1)} \cdots g_{11}^{(1)} \cdot g_2^{(1)}. \tag{3.52a}$$

$$E\left(c_{00}\, b_{s_2 0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} P_{s_2}\, f \cdot G \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\; g_{m-2,1}$$
$$\cdots g_{11} \cdot g_2^{(1)}. \tag{3.52b}$$

$$E\left(c_{s_1 t}\, b_{s_2 0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} P_{s_2} f \cdot G_{s_1 t} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_{s_1+1}=0}^{\tilde{p}_{s_1+1}} \sum_{i_{s_1}=0}^{\tilde{p}_{s_1+1}} \sum_{i_{s_1-1}=0}^{\tilde{p}_{s_1-1}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\; g_{m-2,1} \cdots g_{s_1+1,1}\; g^{(2)}_{s_1,1}\; g_{s_1-1,1} \cdots g_{11} \cdot g_2. \tag{3.52c}$$

$$E\left(c_{0t}\, b_{s_2 0} \mid O^{1,n}\right) = F_m \sum_{s^{1,n}\in\Omega_n} P_{s_2}\, f \cdot G_{0t} \sum_{i_{m-1}=0}^{\tilde{p}_{m-1}} \sum_{i_{m-2}=0}^{\tilde{p}_{m-2}} \cdots \sum_{i_1=0}^{\tilde{p}_1} g_{m-1,1}\; g_{m-2,1} \cdots g_{11} \cdot g_2^{(2)}. \tag{3.52d}$$

2. LSE with $m_A = 2$, $m_B = 2$

First, we consider the case that the state space for both state and observation sequences is $\{0, 1\}$; i.e., $m_A = m_B = 2$ and so both A and B are 2×2 matrices, while π is a vector of length 2 as shown below.

$$\bar{\pi} = (\pi_1\ \pi_0),$$

$$A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} = \begin{pmatrix} c_{00} & 1-c_{00} \\ 1-c_{10} & c_{10} \end{pmatrix} = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix},$$

and

$$B = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix} = \begin{pmatrix} b_{00} & 1-b_{00} \\ b_{10} & 1-b_{10} \end{pmatrix} = \begin{pmatrix} x & 1-x \\ 1-y & y \end{pmatrix},$$

where

$$\pi_i = P\left(S_1 = i\right),$$

$$a_{ij} = P\left(S_{t+1} = j \mid S_t = i\right), \text{and}$$

$$b_{ik} = P\left(O_t = k \mid S_t = i\right)$$

for $t = 1, 2, \ldots, n$ if the sequence length is n.

2.1. Derivation. Letting the initial probabilities $\pi_0 = \pi_1 = \frac{1}{2}$, we have

$$\theta_{LS} = E\left(\theta \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} \int \theta b_{s_1o_1} a_{s_1s_2} b_{s_2o_2} \cdots a_{s_{n-1}s_n} b_{s_no_n} d\theta \tag{3.53}$$

where

$$P\left(O^{1,n}\right) = \frac{1}{2} \sum_{s^{1,n}\in\Omega_n} \int b_{s_1o_1} a_{s_1s_2} b_{s_2o_2} \cdots a_{s_{n-1}s_n} b_{s_no_n} d\theta. \tag{3.54}$$

But, as for the integration with respect to θ, the symmetry in the HMM has to be taken care of first.

2.1.1. *Symmetry.* The two states in the state sequence, say State k_1 and State k_2, can be represented in two ways; i.e., we can choose 0 to represent State k_1 and 1 to represent State k_2 and vice versa. So, by setting the parameter set (the conditional probabilities) accordingly, we have two different parameter sets that produces the same probability given any observation sequence. Consider two models with parameter sets $\theta = (\pi, A, B)$ and $\bar{\theta} = (\bar{\pi}, \bar{A}, \bar{B})$ such that $\bar{\theta}$ is obtained by interchanging 0's and 1's of the transition in the model using θ. In other words, we define $\bar{\theta}$ as shown below.

Let $\bar{\theta} = (\bar{\pi}, \bar{A}, \bar{B})$ where

$$\bar{\pi} = (\bar{\pi}_0 \bar{\pi}_1) = (\pi_1 \pi_0),$$

$$\bar{A} = \begin{pmatrix} \bar{a} & 1-\bar{a} \\ 1-\bar{b} & \bar{b} \end{pmatrix} = \begin{pmatrix} \bar{a}_{00} & \bar{a}_{01} \\ \bar{a}_{10} & \bar{a}_{11} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{10} \\ a_{01} & a_{00} \end{pmatrix} = \begin{pmatrix} b & 1-b \\ 1-a & a \end{pmatrix},$$

and

$$\bar{B} = \begin{pmatrix} \bar{x} & 1-\bar{x} \\ 1-\bar{y} & \bar{y} \end{pmatrix} = \begin{pmatrix} \bar{b}_{00} & \bar{b}_{01} \\ \bar{b}_{10} & \bar{b}_{11} \end{pmatrix} = \begin{pmatrix} b_{10} & b_{11} \\ b_{00} & b_{01} \end{pmatrix} = \begin{pmatrix} 1-y & y \\ x & 1-x \end{pmatrix}.$$

For example, with $\bar{S}^{1,n}$, which is a sequence obtained by interchanging 0 and 1 in $S^{1,n}$, we have

$$\bar{a}_{01} = P(S_t = 1 \mid S_{t-1} = 0) = P(\bar{S}_t = 0 \mid \bar{S}_{t-1} = 1) = a_{10},$$

and

$$\bar{b}_{11} = P(O_t = 1 \mid S_t = 1) = P(O_t = 1 \mid \bar{S}_t = 0) = b_{01}.$$

Thus, we have

$$P\left(\bar{S}^{1,n}, O^{1,n} \mid \bar{\theta}\right) = P\left(S^{1,n}, O^{1,n} \mid \theta\right).$$

In order to avoid averaging this symmetry in probability out, we evaluate the integrations

$$\int \theta b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \tag{3.55}$$

and

$$\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta, \tag{3.56}$$

which appear in equations (3.53) and (3.54), respectively, under a restriction $a \geq b$.

2.1.2. *Evaluating Integrals.* We first consider the second integration (3.56) under the restriction $a \geq b$. Given a particular set of $s^{1,n}$ and $o^{1,n}$, the expression to integrate is a product of a_{ij} and b_{ij}, $i, j \in \{0, 1\}$. So, we need to count how many times each of these factors are multiplied.

Let k_{ij} be the number of times the event $S_{i-1} = i$ and $S_t = j$ occurs in a sequence $S^{1,n} = (S_1, S_2, \ldots, S_n)$ for $2 \leq t \leq n$, and let l_{ij} be the number of times the event $S_t = i$ and $O_t = j$ occurs with the observation sequence $O^{1,n} = (O_1, O_2, \ldots, O_n)$ for $1 \leq t \leq n$. Then we can rewrite equation (3.56) as

$$\begin{aligned}
&\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \\
&\quad = \int_0^1 \int_0^{a_{00}} \int_0^1 \int_0^1 a_{00}^{k_{00}} (1 - a_{00})^{k_{01}} a_{11}^{k_{11}} (1 - a_{11})^{k_{10}} \\
&\qquad\qquad \cdot b_{00}^{l_{00}} (1 - b_{00})^{l_{01}} b_{10}^{l_{11}} (1 - b_{10})^{l_{10}} \, db_{10} \, db_{00} \, da_{11} \, da_{00} \\
&\quad = \int_0^1 \int_0^{c_{00}} \int_0^1 \int_0^1 c_{00}^{d_{00}} (1 - c_{00})^{d_{01}} c_{10}^{d_{10}} (1 - c_{10})^{d_{11}} \\
&\qquad\qquad \cdot b_{00}^{l_{00}} (1 - b_{00})^{l_{01}} b_{10}^{l_{11}} (1 - b_{10})^{l_{10}} \, db_{10} \, db_{00} \, da_{11} \, da_{00},
\end{aligned} \tag{3.57}$$

where

$$\begin{pmatrix} k_{00} & k_{01} \\ k_{10} & k_{11} \end{pmatrix} = \begin{pmatrix} d_{00} & d_{01} \\ d_{11} & d_{10} \end{pmatrix}.$$

We refer to Section 1, and find the estimates by applying those formulas for $m_A = m$ to $m_A = 2$. (However, since the state space size $m_A = 2$ is small, obviously we also could have find the formula to evaluate the above integral in a much simpler way than shown in Section 1.)

Using the equation (3.17) and the definition of $p_i(j)$, I_k, and f in equations (3.10) and (3.6), we have

$$\begin{aligned} P\left(O^{1,n}\right) &= \frac{1}{2} \sum_{s^{1,n} \in \Omega_n} f \sum_{i_1=0}^{\tilde{p}_1} g_{11}(I_1) g_2(I_1) = \frac{1}{2} \sum_{s^{1,n} \in \Omega_n} f \sum_{i=0}^{d_{11}} g_{11}(i) g_2(i) \\ &= \frac{1}{2} \sum_{s^{1,n} \in \Omega_n} f \sum_{i=0}^{k_{10}} g_{11}(i) g_2(i), \end{aligned}$$

where

$$f = \prod_{i=0}^{1} \frac{l_{i0}!\, l_{i1}!}{(l_{i0} + l_{i1} + 1)!}.$$

Here, by the definitions of g_{11} and g_2 in (3.13),

$$g_{11}(i) = \frac{(-1)^i \tilde{p}_1!}{i!(\tilde{p}_1 - i)!\, \phi_1(i)} = \frac{(-1)^i d_{11}!}{i!(d_{11} - i)!\, (d_{10} + i + 1)} = \frac{(-1)^i k_{10}!}{i!\, (k_{10} - i)!\, (k_{11} + i + 1)}$$

and

$$\begin{aligned} g_2(i) &= \frac{(\phi_1(i) + d_{00})\, \tilde{p}_0!}{(\phi_1(i) + d_{00} + \tilde{p}_0 + 1)!} = \frac{(d_{00} + d_{10} + i + 1)\, d_{01}!}{(d_{00} + d_{01} + d_{10} + i + 2)!} = \frac{(d_{00} + d_{10} + i + 1)\, d_{01}!}{(n - d_{11} + i + 1)!} \\ &= \frac{(k_{00} + k_{11} + i + 1)\, k_{01}!}{(n - k_{10} + i + 1)!}. \end{aligned}$$

We define two equations, using the definitions (3.19) and (3.20) as shown below.

$$g_{11}^{(1)}(i) = \frac{\phi_1(i)}{\phi_1(i) + 1}\, g_{11}(i) = \frac{d_{10} + i + 1}{d_{10} + i + 2}\, g_{11}(i) = \frac{k_{11} + i + 1}{k_{11} + i + 2}\, g_{11}(i)$$

and

$$g_2^{(1)}(i) = \frac{\phi_1(i) + d_{00} + 1}{\phi_1(i) + d_{00} + \tilde{p}_0 + 2}\, g_2(i) = \frac{d_{00} + d_{10} + i + 2}{d_{00} + d_{01} + d_{10} + i + 3}\, g_2(i)$$
$$= \frac{k_{00} + k_{11} + i + 2}{n - k_{10} + i + 2}\, g_2(i)$$

Then, using equations (3.22) and (3.24), and with the definition of P_k in (3.23), we have

$$\hat{a} = \hat{c}_{00} = \frac{1}{2P\,(O^{1,n})} \sum_{s^{1,n} \in \Omega_n} f \sum_{i=0}^{d_{11}} g_{11}(i) g_2^{(1)}(i), \tag{3.58a}$$

$$\hat{b} = \hat{c}_{10} = \frac{1}{2P\,(O^{1,n})} \sum_{s^{1,n} \in \Omega_n} f \sum_{i=0}^{d_{11}} g_{11}^{(1)}(i) g_2^{(1)}(i), \tag{3.58b}$$

$$\hat{x} = \hat{b}_{00} = \frac{1}{2P\,(O^{1,n})} \sum_{s^{1,n} \in \Omega_n} P_0\, f \sum_{i=0}^{d_{11}} g_{11}(i) g_2(i), \qquad \text{and} \tag{3.58c}$$

$$1 - \hat{y} = \hat{b}_{10} = \frac{1}{2P\,(O^{1,n})} \sum_{s^{1,n} \in \Omega_n} P_1\, f \sum_{i=0}^{d_{11}} g_{11}(i) g_2(i), \tag{3.58d}$$

where

$$P_k = \frac{l_{k0} + 1}{l_{k0} + l_{k1} + 2}, \qquad k = 0, 1.$$

2.1.3. *Finding the Exponents.* With $m_A = m_B = 2$, the coefficient matrices $\hat{R}$ and R in equations (3.26) and (3.30), which are used for finding the exponents in the integration, $k_{00}, k_{01}, k_{10}, l_{00}, l_{01}$, and l_{10}, are the same.

$$R = \hat{R} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

So, using the equations (3.41), we express the above exponents as a function of k_{11}, s_1, s_n, l_{11}, and l_1 as shown below.

$$\begin{aligned} d_{00} = k_{00} &= k_{11} - 2k_1 + s_1 + s_n + n - 1, \\ d_{01} = k_{01} &= -k_{11} - s_1 + k_1, \\ d_{11} = k_{10} &= -k_{11} - s_n + k_1, \\ l_{00} &= l_{11} - k_1 - l_1 + n, \\ l_{01} &= -l_{11} + l_1, \qquad \text{and} \\ l_{10} &= -l_{11} + k_1, \end{aligned}$$

where k_1 is the number of 1's in the state sequence $S^{1,n}$.

Define ω as

$$\omega = (k_1, k_{11}, l_{11}, s_1, s_n).$$

2.1.4. *Transition Sequences $s^{1,n}$ that Correspond to the Same ω.* Now we have formulas for the parameter estimators and the method to find the exponents that are used in the formulas; but, as they are the computation is expensive. The summation over all the elements in Ω_n, which is the set of all possible state sequences of length n, means 2^n additions since the state space is size 2. Luckily, many different state sequences have the same corresponding values of $\omega = (k_1, k_{11}, l_{11}, s_1, s_n)$. For example, if $o^{1,7} = (0, 1, 1, 0, 1, 0, 0)$ then there are four different $s^{1,7}$ values that have the same $\omega = (s_1, s_n, k_1, k_{11}, l_{11}) = (0, 1, 3, 1, 1)$, which are

$$s^{1,7} = \begin{cases} (0,1,0,0,0,1,1) \\ (0,0,1,1,0,0,1) \\ (0,0,1,0,0,1,1) \\ (0,0,0,1,1,0,1). \end{cases}$$

Actually, since $s_1, s_n \in \{0,1\}$, $k_1 \in \{0,1,\ldots,n\}$, $k_{11} \in \{0,1,\ldots,n-1\}$, and $l_{11} \in \{0,1,\ldots,\max(k_1,l_1)\}$, we can see the size of $\overline{\Omega}_n$ is less than $4n^3$. Furthermore, since many of the combinations of k_1, k_{11}, s_1, s_n, and l_{11} are impossible given a particular $o^{1,n}$, the size is much smaller than $4n^3$. For example, when $n = 25$, experimentally we see the size of $\overline{\Omega}_n\left(O^{1,n}\right)$ is around $4000 < 2n^4 = 781250$ in average, while the size of Ω_n is $2^n = 2^{25} = 33554432$.

Let $h_n(\omega) = h_n(k_1, k_{11}, l_{11}, s_1, s_n)$ be the number of possible values of $s^{1,n} \in \Omega_n$ given ω and a particular observation sequence $o^{1,n}$.

We observe that a symmetry exists in the distribution of h_n. Let $\bar{s}_t$ be such that

$$\bar{s}_t = \begin{cases} 1 & \text{if } s_t = 0 \\ 0 & \text{if } s_t = 1. \end{cases}$$

Also, let $\bar{\omega}$ be the values you obtain by interchanging the state values in the transition, given ω; i.e.,

$$\bar{\omega} = (k_0, k_{00}, l_{00}, \bar{s}_1, \bar{s}_n) \quad \text{if } \omega = (k_1, k_{11}, l_{11}, s_1, s_n).$$

Because of the symmetry, which comes from the summation over all the possible values of $S^{1,n} \in \Omega_n$, we have

$$h_n(\omega) = h_n(\bar{\omega}).$$

If we fix the value of s_1, say $s_1 = 0$, then it is not possible for the algorithm to consider both ω and its corresponding $\bar{\omega}$ for any value of ω. So, we fix s_1 as 0; i.e., we will first find all the values of $h_n(\omega)$ such that $\omega = (k_1, k_{11}, l_{11}, 0, s_n)$, then assign the same values to the corresponding $h_n(\bar{\omega})$.

Now, as for the algorithm to find h_n (with $s_1 = 0$ fixed), suppose we already have the values of $h_t(\omega)$ for all ω, corresponding to all the values of $s^{1,t}$ and a partial observation sequence $o^{1,t}$. Then, in order to obtain h_{t+1}, there are two cases to consider.

- If the next transition state s_{t+1} is 0, then
  - there will be no change in the values of s_1 (the first state in $s^{1,t+1}$), k_1 (the number of 1's in s^{t+1}), k_{11} (the number of times state transfers from 1 to 1 in $s^{1,t+1}$), and l_{11} (the number of times 1 is observed when the state is 1, given $s^{1,t+1}$ and $o^{1,t+1}$), and
  - only possible change in ω is that the last state s_n will be 0.
- If the next transition state s_{t+1} is 1, then
  - there will be no change in the value of s_1 (the first state in $s^{1,t+1}$),
  - k_1 (the number of 1's in s^{t+1}) will be increased by 1,
  - k_{11} (the number of times state transfers from 1 to 1 in $s^{1,t+1}$) will be increased by 1 if the current state s_t is also 1, and will stay the same otherwise,
  - l_{11} (the number of times 1 is observed when the state is 1, given $s^{1,t+1}$ and $o^{1,t+1}$) will be increased by 1 if the next emission state o_{t+1} is 1, and will stay the same otherwise, and
  - the last state s_n will be 1.

So, we start from one observed state, for which the values of h_1 are obvious, and keep processing the next observation state till we find h_n. The result is the algorithm for $h_n(\omega) = h_n(s_1, s_n, k_1, k_{11}, l_{11})$ shown below. Note that while it go through the for-loop, it also goes through the values of ω, but only through those that is possible to obtain given $o^{1,t}$.

Algorithm for finding h_n:

Let $h_1(0,0,0,0,0) = 1$.

for t from 1 to $n-1$

with all $\omega = (k_1, k_{11}, l_{11}, s_1, s_t)$ such that $h_t(\omega) > 0$

increment $h_{t+1}(\ k_1,\ k_{11},\ l_{11},\ s_1,\ 0)$ and $h_{t+1}(k_1+1,\ k_{11}+s_t,\ l_{11}+o_{t+1},\ s_1,\ 1)$

by the value $h_t(\omega)$

end for

Since the size of $\overline{\Omega}_t$ is much less than $4t^3$ for each $t = 1, 2, \ldots, n$, and the for-loop repeats $n-1$ times, the number of addition is less than cn^4 for some constant c.

Figure 1 shows an experimental result for the size of $\overline{\Omega}_n$ for $n = 40$. Three hundred $\theta = (a, b, x, y, r)$ values are randomly selected, then for each θ, a state sequence and an observation sequence of length 40 are generated. The mean of the size of $\overline{\Omega}_{40}$, or the number of distinguishable ω-values, was 6525.81 which is considerably smaller than the size of Ω_{40}, which is $2^{40} \approx 1.1^{12}$.

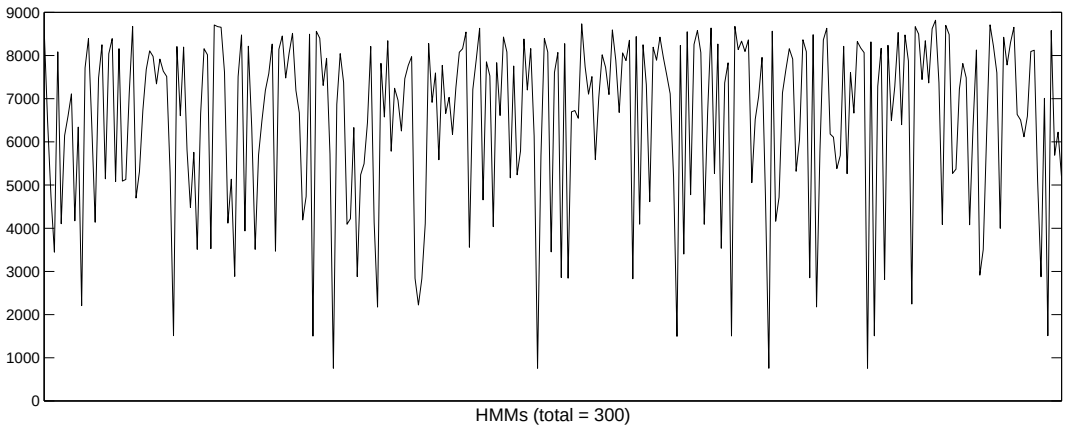


Figure 1. Change in the size of $\overline{\Omega}_{40}$ with $m_A = 2$ and sequence length 40, over 300 randomly generated observation sequences.

2.2. Covariance Matrix. Here we simply use the formula for estimates obtained above, while referring to the formulas shown in Section 1.2, to find the formulas for the covariance matrix.

Using the definitions (3.47), let

$$g_{11}^{(11)}(i) = \frac{\phi_1(i)}{\phi_1(i)+2}\, g_{11}(i) = \frac{d_{10}+i+1}{d_{10}+i+3}\, g_{11}(i) = \frac{k_{11}+i+1}{k_{11}+i+3}\, g_{11}(i),$$

and

$$\begin{aligned} g_2^{(11)}(i) &= \frac{\phi_1(i)+d_{00}+2}{\phi_1(i)+d_{00}+\tilde{p}_0+3}\, g_2^{(1)}(i) = \frac{d_{00}+d_{10}+i+3}{d_{00}+d_{10}+i+d_{01}+4}\, g_2^{(1)}(i) \\ &= \frac{k_{00}+k_{11}+i+3}{n-k_{10}+i+3}\, g_2^{(1)}(i). \end{aligned}$$

Then, by equations (3.49), we have

$$E\left(aa \mid O^{1,n}\right) = E\left(c_{00}c_{00} \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \sum_{i=0}^{k_{10}} g_{11}(i)\, g_2^{(11)}(i),$$

$$E\left(bb \mid O^{1,n}\right) = E\left(c_{10}c_{10} \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \sum_{i=0}^{k_{10}} g_{11}^{(11)}(i)\, g_2^{(11)}(i), \text{ and}$$

$$E\left(ab \mid O^{1,n}\right) = E\left(c_{00}c_{10} \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \sum_{i=0}^{k_{10}} g_{11}^{(1)}(i)\, g_2^{(11)}(i).$$

Also, by equations (3.51), we have

$$E\left(b_{k0}b_{k0} \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} \widetilde{P}_k P_k f \sum_{i=0}^{k_{10}} g_{11}(i)\, g_2(i)$$

and

$$E\left(b_{00}b_{10} \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_0 P_1 f \sum_{i=0}^{k_{10}} g_{11}(i)\, g_2(i)$$

for $k = 0, 1$ where

$$\widetilde{P}_k = \frac{l_{k0} + 2}{l_{k0} + l_{k1} + 3}.$$

Furthermore, by the equations (3.52), we have

$$E\left(c_{00}b_{k0} \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_k f \sum_{i=0}^{k_{10}} g_{11}(i)\, g_2^{(1)}(i) \text{ and}$$

$$E\left(c_{10}b_{k0} \mid O^{1,n}\right) = \frac{1}{2P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_k f \sum_{i=0}^{k_{10}} g_{11}^{(1)}(i)\, g_2^{(1)}(i)$$

for $k = 0, 1$.

3. LSE with $m_A = 3$, $m_B = 2$

We consider here the case that the state space is $\{0, 1, 2\}$ for the state sequence and $\{0, 1\}$ for the observation sequence; i.e., $m_A = 3$ and $m_B = 2$. Let the transition matrix A

and the emission matrix B be as shown below.

$$A = \begin{pmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} c_{00} & c_{01} & 1 - c_{00} - c_{01} \\ 1 - c_{10} - c_{11} & c_{10} & c_{11} \\ c_{21} & 1 - c_{20} - c_{21} & c_{20} \end{pmatrix}$$

$$B = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \\ b_{20} & b_{21} \end{pmatrix} = \begin{pmatrix} b_{00} & 1 - b_{00} \\ b_{10} & 1 - b_{10} \\ b_{20} & 1 - b_{20} \end{pmatrix}$$

The parameters we actually estimate are now, c_{ij} and b_{i0} for $i \in \{0, 1, 2\}$ and $j \in \{0, 1\}$.

3.1. Derivation. Letting the initial probabilities $\pi_0 = \pi_1 = \pi_2 = \frac{1}{3}$ in equations (3.1) and (3.2), we have

$$\theta_{LS} = E\left(\theta \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} \int \theta b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \tag{3.59}$$

where

$$P\left(O^{1,n}\right) = \frac{1}{3} \sum_{s^{1,n} \in \Omega_n} \int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta. \tag{3.60}$$

Note that Ω_n is now all the possible length-n state sequence of 0, 1, and 2. Thus, the size of Ω_n is 3^n. Also, as same as for the case with $m_A = m_B = 2$, we need to consider the symmetry in the HMM first to compute the integrals.

3.1.1. *Symmetry.* Similar to the case when the state space for the state sequence is $\{0, 1\}$, we could identify three states in the state sequence in six different ways with $\{0, 1, 2\}$ and so have symmetries in the probability distribution. Hence, we integrate under the restriction that $a_{00} \geq a_{11} \geq a_{22}$, which is same as $c_{00} \geq c_{10} \geq c_{20}$.

3.1.2. *Evaluating Integrals.* As for the integrals in equations (3.59) and (3.60), we use the result from the previous section, Section 1. As before, let k_{ij} be the number of times the event $S_{t-1} = i$ and $S_t = j$ occurs in a sequence $S^{1,n} = (S_1, S_2, \ldots, S_n)$ for $2 \le t \le n$, and let l_{ij} be the number of times the event $S_t = i$ and $O_t = j$ occurs with the observation sequence $O^{1,n} = (O_1, O_2, \ldots, O_n)$ for $1 \le t \le n$. Also, let $K = \{k_{ij}\}$ and $L = \{l_{iu}\}$, $i, j \in \{0, 1, 2\}$, $u \in \{0, 1\}$, denote the set of those values.

Furthermore, we rewrite the exponents k_{ij} with d_{st} as shown below.

$$\begin{pmatrix} k_{00} & k_{01} & k_{02} \\ k_{10} & k_{11} & k_{12} \\ k_{20} & k_{21} & k_{22} \end{pmatrix} = \begin{pmatrix} d_{00} & d_{01} & d_{02} \\ d_{12} & d_{10} & d_{11} \\ d_{21} & d_{22} & d_{20} \end{pmatrix} \tag{3.61}$$

Using the equations (3.15) and the definitions (3.10) and (3.13), we have

$$\begin{aligned} &\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \\ &\qquad = f \cdot \prod_{k=0}^{2} \sum_{i=0}^{d_{k2}} g_{k0}(i) \sum_{i_2=0}^{\tilde{p}_2} g_{21}(i_2) \sum_{i_1=0}^{\tilde{p}_1} g_{11}(i_1, i_2)\, g_2(i_1, i_2) \end{aligned} \tag{3.62}$$

where

$$\begin{aligned} f &= \prod_{i=0}^{2} \frac{l_{i0}! l_{i1}!}{(l_{i0} + l_{i1} + 1)!}, \\ g_{k0}(i) &= \frac{(-1)^i d_{k2}!}{i!\,(d_{k2} - i)!\,(d_{k1} + i + 1)}, \\ g_{21}(i_2) &= \frac{(-1)^{i_2}\, \tilde{p}_2!}{i_2!(\tilde{p}_2 - i_2)!\phi_2(i_2)}, \\ g_{11}(i_1, i_2) &= \frac{(-1)^{i_1} \tilde{p}_1!}{i_1!(\tilde{p}_1 - i_1)!\phi_1(i_1, i_2)}, \\ g_2(i_1, i_2) &= \frac{(\phi_1(i_1, i_2) + d_{00})!\, \tilde{p}_0!}{(\phi_1(i_1, i_2) + d_{00} + \tilde{p}_0 + 1)!}, \end{aligned}$$

$$\tilde{p}_i = d_{i1} + d_{i2} + 1,$$

$$\phi_1(i_1, i_2) = \phi_2(i_2) + d_{10} + i_1 + 1, \qquad \text{and}$$

$$\phi_2(i_2) = d_{20} + i_2 + 1. \tag{3.63}$$

Thus, we rewrite the equation (3.60) as

$$\begin{aligned} P\left(O^{1,n}\right) &= \frac{1}{3} \sum_{s^{1,n} \in \Omega_n} \int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \\ &= \frac{1}{3} \sum_{s^{1,n} \in \Omega_n} f \cdot \prod_{k=0}^{2} \sum_{i=0}^{d_{k2}} g_{k0}(i) \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}(i_2)\; g_{11}(i_1, i_2)\; g_2(i_1, i_2). \end{aligned} \tag{3.64}$$

Also, we define the following equations, using the definitions (3.18), (3.19), and (3.20):

$$g_{k0}^{(1)}(i) = \frac{d_{k1} + i + 1}{d_{k1} + i + 2}\; g_{k0}(i), \tag{3.65a}$$

$$g_{11}^{(1)}(i_1, i_2) = \frac{\phi_1(i_1, i_2)}{\phi_1(i_1, i_2) + 1}\; g_{11}(i_1, i_2), \tag{3.65b}$$

$$g_{11}^{(2)}(i_1, i_2) = \frac{(-1)^{i_1} (\tilde{p}_1 + 1)!}{i_1! (\tilde{p}_1 - i_1 + 1)! \phi_1(i_1, i_2)}, \tag{3.65c}$$

$$g_{21}^{(1)}(i_2) = \frac{\phi_2(i_2)}{\phi_2(i_2) + 1}\; g_{21}(i_2), \tag{3.65d}$$

$$g_{21}^{(2)}(i_2) = \frac{(-1)^{i_2}\, (\tilde{p}_2 + 1)!}{i_2! (\tilde{p}_2 - i_2 + 1)! \phi_2(i_2)}, \tag{3.65e}$$

$$g_2^{(1)}(i_1, i_2) = \frac{\phi_1(i_1, i_2) + d_{00} + 1}{\phi_1(i_1, i_2) + d_{00} + \tilde{p}_0 + 2}\; g_2(i_1, i_2) \quad \text{and} \tag{3.65f}$$

$$g_2^{(2)}(i_1, i_2) = \frac{\tilde{p}_0 + 1}{\phi_1(i_1, i_2) + d_{00} + \tilde{p}_0 + 2}\; g_2(i_1, i_2). \tag{3.65g}$$

Using the definitions (3.16) and (3.21) for G and G_{st}, respectively, we get the following estimates from the equations (3.22):

$$\hat{c}_{00} = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}(i_2)\; g_{11}(i_1, i_2)\; g_2^{(1)}(i_1, i_2,) \tag{3.66a}$$

$$\hat{c}_{10} = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}(i_2)\; g_{11}^{(1)}\;(i_1,i_2)\; g_2^{(1)}(i_1,i_2), \tag{3.66b}$$

$$\hat{c}_{20} = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(1)}(i_2)\; g_{11}^{(1)}(i_1,i_2)\; g_2^{(1)}(i_1,i_2), \tag{3.66c}$$

$$\hat{c}_{01} = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_0 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}(i_2)\; g_{11}(i_1,i_2)\; g_2^{(2)}(i_1,i_2), \tag{3.66d}$$

$$\hat{c}_{11} = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_1 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1+1} g_{21}(i_2)\; g_{11}^{(2)}(i_1,i_2)\; g_2(i_1,i_2), \tag{3.66e}$$

$$\hat{c}_{21} = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_2 \sum_{i_2=0}^{\tilde{p}_2+1} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(2)}(i_2)\; g_{11}(i_1,i_2)\; g_2(i_1,i_2), \tag{3.66f}$$

and

$$\hat{b}_{k0} = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_k \cdot f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}(i_2)\; g_{11}(i_1,i_2)\; g_2(i_1,i_2), \tag{3.66g}$$

where

$$G = \prod_{k=0}^{2} \sum_{i=0}^{d_{k2}} g_{k0}(i) \tag{3.67}$$

and

$$G_s = G_{s1} = \sum_{i=0}^{d_{s2}} g_{s0}^{(1)}(i) \cdot \prod_{\substack{k=0 \\ k\neq s}}^{2} \sum_{i=0}^{d_{k2}} g_{k0}(i) \qquad \text{for } s \in \{0,1,2\}.$$

Also, using the notation G, we can express $P\left(O^{1,n}\right)$ as

$$P\left(O^{1,n}\right) = \frac{1}{3} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}(i_2)\; g_{11}(i_1,i_2)\; g_2(i_1,i_2).$$

3.1.3. *Finding Exponents.* We try to rewrite exponent sets K and L so that they are expressed as sums and/or differences of quantities that are eventually converted to quantities that will be used in the algorithm to find the exponents efficiently.

Define δ and γ as a function of s_t and o_t, respectively, as

$$\delta_i(j) = \begin{cases} 1 & \text{if } i = j \\ \\ 0 & \text{if } i \neq j \end{cases} \qquad \text{and} \qquad \gamma_u(v) = \begin{cases} 1 & \text{if } u = v \\ \\ 0 & \text{if } u \neq v \end{cases}$$

where $i, j \in \{0, 1, 2\}$ and $u, v \in \{0, 1\}$.

Let R and M be as defined in Section 1. Then, we have

$$M = \begin{pmatrix} 1 & 1 & 1 \\ \\ 0 & 1 & 2 \\ \\ 0 & 1 & 2^2 \end{pmatrix}$$

so that

$$R = M^{-1} = \begin{pmatrix} r_{00} & r_{01} & r_{02} \\ \\ r_{10} & r_{11} & r_{12} \\ \\ r_{20} & r_{21} & r_{22} \end{pmatrix} = \begin{pmatrix} 1 & -\frac{3}{2} & \frac{1}{2} \\ \\ 0 & 2 & -1 \\ \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

and

$$\delta_i(j) = \begin{cases} \frac{1}{2}\left(2 - 3j + j^{\,2}\right) & \text{if } i = 0 \\ \\ 2j - j^{\,2} & \text{if } i = 1 \\ \\ \frac{1}{2}\left(-j + j^{\,2}\right) & \text{if } i = 2. \end{cases}$$

We will find K and L using the values of r_{ij} above and the quantities below.

$$k_1 = \text{number of 1's in } s^{1,n}$$

$$k_2 = \text{number of 2's in } s^{1,n}$$

$$k_{11} = \text{the number of 11's in } s^{1,n}$$

$$k_{12} = \text{the number of 12's in } s^{1,n}$$

$$k_{21} = \text{the number of 21's in } s^{1,n}$$

$$k_{22} = \text{the number of 22's in } s^{1,n}$$

$$l_1 = \text{number of 1's in } o^{1,n}$$

$$l_{11} = \text{number of times } s_t = 1 \text{ and } o_t = 1, \text{ for } 1 \leq t \leq n$$

$$l_{21} = \text{number of times } s_t = 2 \text{ and } o_t = 1, \text{ for } 1 \leq t \leq n$$

Then, using the results from Section 1, we have the following:

$$k_{00} = k_{22} + k_{21} + k_{12} + k_{11} - 2k_2 - 2k_1 + \frac{1}{2}\left(-s_1^2 - s_n^2 + 3s_1 + 3s_n\right) + n - 1$$

$$k_{01} = -k_{21} - k_{11} + k_1 + s_1^2 - 2s_1$$

$$k_{02} = -k_{22} - k_{12} + k_2 + \frac{1}{2}\left(-s_1^2 + s_1\right)$$

$$k_{10} = -k_{12} - k_{11} + k_1 + s_n^2 - 2s_n$$

$$k_{20} = -k_{21} - k_{22} + k_2 + \frac{1}{2}\left(-s_n^2 + s_n\right)$$

and

$$l_{00} = l_{11} + l_{21} - k_2 - k_1 - l_1 + n$$

$$l_{01} = -l_{11} - l_{21} + l_1$$

$$l_{10} = -l_{11} + k_1$$

$$l_{20} = -l_{21} + k_2.$$

Now, L and K can be expressed as a function of ω where

$$\omega = (k_1, k_2, k_{11}, k_{12}, k_{21}, k_{22}, l_{11}, l_{21}, s_1, s_n)$$

As before, let $h_n(\omega) = h_n(k_1, k_2, k_{11}, k_{12}, k_{21}, k_{22}, l_{11}, l_{21}, s_1, s_n)$ be the number of possible values of $s^{1,n} \in \Omega_n$ given a particular observation sequence $o^{1,n}$. Then, for example $P\left(O^{1,n}\right)$ can be computed by summing over $\overline{\Omega}_n$, which is the set of all ω-values that corresponds to the combination of some $s^{1,n} \in \Omega_n$ and a particular observation sequence $o^{1,n}$.

We find the values of h_n in a similar method used for the case $m_A = m_B = 2$ described in Section 2. Suppose we already have the values for $h_t(\omega)$ for all ω, corresponding all the values of $s^{1,t}$ and a partial observation sequence $o^{1,t}$. Then, in order to process the next observation o_{t+1}, there are three cases to consider.

- If the next transition state s_{t+1} is 0, then
  - the last state s_n will be 0, and
  - the rest of the quantities will stay the same
- For $j \in \{1, 2\}$, if the next transition state s_{t+1} is j, then
  - the last state s_n will be j,
  - k_j (the number of j's in $s^{1,t+1}$) will be increased by 1,
  - k_{ij} (the number of ij's in $s^{1,t+1}$) will be increased by 1 if the current state s_t is i, and will stay the same otherwise, for $i \in \{1, 2\}$,
  - l_{j1} (the number of times 1 is observed when the state is j, given $s^{1,t+1}$ and $o^{1,t+1}$) will be increased by 1 if $o_{t+1} = 1$, and stays the same otherwise, and
  - the rest of the quantities will stay the same.

As before, we start from one observed state o_1, find h_1, then iteratively find $h_{t+1}(\omega)$ till $t+1=n$. Also, there are symmetries with the values of h_n that correspond to ω that has $s_1=0$, $s_1=1$, and $s_1=2$. So, we need just to find the h_n-values that correspond to the set of ω-values for which, say, $s_1=0$ is fixed.

Algorithm for finding h_n:

Let $h_1(0,0,0,0,0,0,0,0,0,0)=1$.

for t from 1 to $n-1$

with all $\omega=(k_1,\, k_2,\, k_{11},\, k_{12},\, k_{21},\, k_{22},\, l_{11},\, l_{21}\, 0,\, s_t)$ such that $h_t(\omega)>0$

Increment $h_{t+1}(k_1,\, k_2,\, k_{11},\, k_{12},\, k_{21},\, k_{22},\, l_{11},\, l_{21},\, 0,\, 0)$ by the value of $h_t(\omega)$.

for $j=1$ and 2

Obtain $\hat{\omega}$ from ω by incrementing k_j by one, $k_{s_t,\,j}$ by one, and l_{j1} by o_{t+1}; then replacing s_t by j.

Increment $h_{t+1}(\hat{\omega})$ by the value of $h_t(\omega)$.

end for

end for

Since the algorithm goes through the for-loop $n-1$ times, and the size of $\overline{\Omega}_n$ is less than n^8, the total number of addition is less than cn^k where $k=m_A^2=3^2=9$ for some constant c.

3.2. Covariance Matrix. As for the covariance matrix, we want to find $E\left(c_{s_1t_1}\, c_{s_2t_2} \mid O^{1,n}\right)$, $E\left(b_{s_10}\, b_{s_20} \mid O^{1,n}\right)$, and $E\left(c_{s_1t}\, b_{s_20} \mid O^{1,n}\right)$.

First, using the definitions (3.47) in Section 1.2, we define new functions as shown below.

$$g_{k0}^{(11)}(i) = \frac{d_{k1}+i+1}{d_{k1}+i+3}\, g_{k0}(i), \tag{3.68a}$$

$$g_{11}^{(11)}(i_1,\, i_2) = \frac{\phi_1(i_1,\, i_2)}{\phi_1(i_1,\, i_2)+2}\, g_{11}(i_1,\, i_2), \tag{3.68b}$$

$$g_{11}^{(22)}(i_1,\, i_2) = \frac{(-1)^{i_1}(\tilde{p}_1+2)!}{i_1!(\tilde{p}_1-i_1+2)!\phi_1(i_1,\, i_2)}, \tag{3.68c}$$

$$g_{11}^{(12)}(i_1,\, i_2) = \frac{(-1)^{i_1}(\tilde{p}_1+1)!}{i_1!(\tilde{p}_1-i_1+1)!\,(\phi_1(i_1,\, i_2)+1)}, \tag{3.68d}$$

$$g_{21}^{(11)}(i_2) = \frac{\phi_2(i_2)}{\phi_2(i_2)+2}\, g_{21}(i_2), \tag{3.68e}$$

$$g_{21}^{(22)}(i_2) = \frac{(-1)^{i_2}(\tilde{p}_2+2)!}{i_2!(\tilde{p}_2-i_2+2)!\phi_2(i_2)}, \tag{3.68f}$$

$$g_{21}^{(12)}(i_2) = \frac{(-1)^{i_2}(\tilde{p}_2+1)!}{i_2!(\tilde{p}_2-i_2+1)!\,(\phi_2(i_2)+1)}, \tag{3.68g}$$

$$g_{2}^{(11)}(i_1,\, i_2) = \frac{\phi_1(i_1,\, i_2)+d_{00}+2}{\phi_1(i_1,\, i_2)+d_{00}+\tilde{p}_0+3}\, g_{2}^{(1)}(i_1,\, i_2), \tag{3.68h}$$

$$g_{2}^{(22)}(i_1,\, i_2) = \frac{\tilde{p}_0+2}{\phi_1(i_1,\, i_2)+d_{00}+\tilde{p}_0+3}\, g_{2}^{(2)}(i_1,\, i_2), \quad \text{and} \tag{3.68i}$$

$$g_{2}^{(12)}(i_1,\, i_2) = \frac{\tilde{p}_0+1}{\phi_1(i_1,\, i_2)+d_{00}+\tilde{p}_0+3}\, g_{2}^{(1)}(i_1,\, i_2). \tag{3.68j}$$

Also, using the equation (3.48) with $s_1, s_2 \in \{0,1,2\}$, we define $G_{s_1\, s_2}$ as follows:

If $s_1 = s_2 = s$,

$$G_{ss} = \sum_{i=0}^{d_{s2}} g_{s0}^{(11)}(i) \cdot \prod_{\substack{k=0 \\ k\neq s}}^{2} \sum_{i=0}^{d_{k2}} g_{k0}, \tag{3.69a}$$

and if $s_1 \neq s_2$ then

$$G_{s_1\, s_2} = \sum_{i=0}^{d_{s_1,\,2}} g_{s_1,\,0}^{(1)}(i) \cdot \sum_{i=0}^{d_{s_2,\,2}} g_{s_2,\,0}^{(1)}(i) \cdot \prod_{\substack{k=0 \\ k\neq s_1,s_2}}^{2} \sum_{i=0}^{d_{k2}} g_{k0}. \tag{3.69b}$$

Using the equations (3.49), (3.51), and (3.52), we can write the expected values in the form $E\left(uv \mid O^{1,n}\right)$ as follows (while, as in the general case $m_A = m$, the indices for g_{21}, g_{11}, and g_2 are fixed and therefore omitted below):

$$E\left(c_{00}c_{00} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\, g_{11}\, g_2^{(11)} \tag{3.70a}$$

$$E\left(c_{00}c_{01} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_0 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\, g_{11}\, g_2^{(12)} \tag{3.70b}$$

$$E\left(c_{00}c_{10} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\, g_{11}^{(1)}\, g_2^{(11)} \tag{3.70c}$$

$$E\left(c_{00}c_{20} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(1)}\, g_{11}^{(1)}\, (g_2^{(11)} \tag{3.70d}$$

$$E\left(c_{00}c_{11} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_1 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1+1} g_{21}\, g_{11}^{(2)}\, g_2^{(1)} \tag{3.70e}$$

$$E\left(c_{00}c_{21} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_2 \sum_{i_2=0}^{\tilde{p}_2+1} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(2)}\, g_{11}\, g_2^{(1)} \tag{3.70f}$$

$$E\left(c_{01}c_{01} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{00} \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\, g_{11}\, g_2^{(22)} \tag{3.70g}$$

$$E\left(c_{01}c_{10} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_0 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\, g_{11}^{(1)}\, g_2^{(12)} \tag{3.70h}$$

$$E\left(c_{01}c_{20} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_0 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(1)}\, g_{11}^{(1)}\, g_2^{(12)} \tag{3.70i}$$

$$E\left(c_{01}c_{11} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{10} \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1+1} g_{21}\, g_{11}^{(2)}\, g_2^{(2)} \tag{3.70j}$$

$$E\left(c_{01}c_{21} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{20} \sum_{i_2=0}^{\tilde{p}_2+1} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(2)}\, g_{11}\, g_2^{(2)} \tag{3.70k}$$

$$E\left(c_{10}c_{10} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2(0)} \sum_{i_1=0}^{\tilde{p}_1(0)} g_{21}\, g_{11}^{(11)}\, g_2^{(11)} \tag{3.70l}$$

$$E\left(c_{20}c_{20} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(11)}\, g_{11}^{(11)}\, g_2^{(11)} \tag{3.70m}$$

$$E\left(c_{10}c_{20} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(1)}\, g_{11}^{(11)}\, g_2^{(11)} \tag{3.70n}$$

$$E\left(c_{10}c_{11} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_1 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1+1} g_{21}\, g_{11}^{(12)}\, g_2^{(1)} \tag{3.70o}$$

$$E\left(c_{10}c_{21} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_2 \sum_{i_2=0}^{\tilde{p}_2+1} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(2)}\, g_{11}^{(1)}\, g_2^{(1)} \tag{3.70p}$$

$$E\left(c_{20}c_{11} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_1 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1+1} g_{21}^{(1)}\, g_{11}^{(12)}\, g_2^{(1)} \tag{3.70q}$$

$$E\left(c_{20}c_{21} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_2 \sum_{i_2=0}^{\tilde{p}_2+1} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(12)}\, g_{11}^{(1)}\, g_2^{(1)} \tag{3.70r}$$

$$E\left(c_{11}c_{11} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{11} \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1+2} g_{21}\, g_{11}^{(22)}\, g_2 \tag{3.70s}$$

$$E\left(c_{21}c_{21} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{22} \sum_{i_2=0}^{\tilde{p}_2+2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(22)}\, g_{11}\, g_2 \tag{3.70t}$$

$$E\left(c_{11}c_{21} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{12} \sum_{i_2=0}^{\tilde{p}_2+1} \sum_{i_1=0}^{\tilde{p}_1+1} g_{21}^{(2)}\, g_{11}^{(2)}\, g_2 \tag{3.70u}$$

As for the expected values $E\left(b_{s_1,0}b_{s_2,0} \mid O^{1,n}\right)$, $s_1, s_2 \in \{0,1,2\}$,we define $\widetilde{P}_s$ as

$$\widetilde{P}_s = \frac{l_{s0}+2}{l_{s0}+l_{s1}+3} \tag{3.71}$$

as previously defined in equation (3.50). Then, if $s_1 = s_2 = s$, we have

$$E\left(b_{s0}b_{s0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} \widetilde{P}_s P_s \cdot f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\, g_{11}\, g_2, \tag{3.72a}$$

and if otherwise, then

$$E\left(b_{s_1 0} b_{s_2 0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_{s_1} P_{s_2} \cdot f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\ g_{11}\ g_2. \tag{3.72b}$$

The expected values in the form $E\left(c_{s_1 t} b_{s_2 0} \mid O^{1,n}\right)$ are as shown below.

$$E\left(c_{10} b_{s0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_s f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\ g_{11}^{(1)}\ g_2^{(1)} \tag{3.73a}$$

$$E\left(c_{20} b_{s0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_s f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(1)}\ g_{11}^{(1)}\ g_2^{(1)} \tag{3.73b}$$

$$E\left(c_{00} b_{s0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_s f \cdot G \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\ g_{11}\ g_2^{(1)} \tag{3.73c}$$

$$E\left(c_{11} b_{s0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_s f \cdot G_1 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1+1} g_{21}\ g_{11}^{(2)}\ g_2 \tag{3.73d}$$

$$E\left(c_{21} b_{s0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_s f \cdot G_2 \sum_{i_2=0}^{\tilde{p}_2+1} \sum_{i_1=0}^{\tilde{p}_1} g_{21}^{(2)}\ g_{11}\ g_2 \tag{3.73e}$$

$$E\left(c_{01} b_{s0} \mid O^{1,n}\right) = \frac{1}{3P\left(O^{1,n}\right)} \sum_{s^{1,n} \in \Omega_n} P_s f \cdot G_0 \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{21}\ g_{11}\ g_2^{(2)} \tag{3.73f}$$

4. LSE with $m_A = 5$, $m_B = 2$

Here we consider the state sequence with state space size five. The transition matrix A and the emission matrix B are now

$$A = \begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} & a_{04} \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} \\ a_{30} & a_{31} & a_{32} & a_{33} & a_{34} \\ a_{40} & a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} c_{00} & c_{01} & c_{02} & c_{03} & c_{04} \\ c_{14} & c_{10} & c_{11} & c_{12} & c_{13} \\ c_{23} & c_{24} & c_{20} & c_{21} & c_{22} \\ c_{32} & c_{33} & c_{34} & c_{30} & c_{31} \\ c_{41} & c_{42} & c_{43} & c_{44} & c_{40} \end{pmatrix}$$

and

$$B = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \\ b_{20} & b_{21} \\ b_{30} & b_{31} \\ b_{40} & b_{41} \end{pmatrix}.$$

Furthermore, as in Section 1 we use the expression d_{ij} for the exponent that corresponds to the parameter c_{ij}, which appears in the integration (3.75) below. So, we have

$$\begin{pmatrix} k_{00} & k_{01} & k_{02} & k_{03} & k_{04} \\ k_{10} & k_{11} & k_{12} & k_{13} & k_{14} \\ k_{20} & k_{21} & k_{22} & k_{23} & k_{24} \\ k_{30} & k_{31} & k_{32} & k_{33} & k_{34} \\ k_{40} & k_{41} & k_{42} & k_{43} & k_{44} \end{pmatrix} = \begin{pmatrix} d_{00} & d_{01} & d_{02} & d_{03} & d_{04} \\ d_{14} & d_{10} & d_{11} & d_{12} & d_{13} \\ d_{23} & d_{24} & d_{20} & d_{21} & d_{22} \\ d_{32} & d_{33} & d_{34} & d_{30} & d_{31} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{40} \end{pmatrix}. \tag{3.74}$$

4.1. Derivation. We use the same definition as for the exponents k_{ij} and l_{ij}; i.e., we let k_{ij} be the number of times the event $S_{i-1} = i$ and $S_t = j$ occurs in a state sequence $S^{1,n} = (S_1, S_2, \ldots, S_n)$ for $2 \le t \le n$, and let l_{ij} be the number of times the event $S_t = i$ and $O_t = j$ occurs with the observation sequence $O^{1,n} = (O_1, O_2, \ldots, O_n)$ for $1 \le t \le n$. The formula for exact LSE can be derived using the results from Section 1, which describes the method for the general case $m_A = m$ for any positive number m.

From the equation (3.15) and the definitions (3.10) in Section 1, with $m = 5$ we have

$$\begin{aligned} &\int b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{n-1} s_n} b_{s_n o_n} d\theta \\ &\quad = f \cdot \prod_{k=0}^{4} \left[\prod_{j=-1}^{1} \sum_{i=0}^{p_k(j)} g_{k0}(i,j) \right] \cdot \sum_{i_4=0}^{\tilde{p}_4} \sum_{i_3=0}^{\tilde{p}_3} \sum_{i_2=0}^{\tilde{p}_2} \sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\, g_{31}(I_3)\, g_{21}(I_2)\, g_{11}(I_1)\, g_2(I_1), \end{aligned} \tag{3.75}$$

where

$$\begin{cases} p_k(-1) = d_{k4} \\ p_k(i) = p_k(i-1) + d_{k,\,3-i} + 1 \quad \text{for } i = 0, 1 \end{cases} \tag{3.76a}$$

$$\tilde{p}_k = p_k(2) = p_k(1) + d_{k,\,1} + 1 \tag{3.76b}$$

$$\begin{cases} \phi_k(I_k) = \phi_{k+1}(I_{k+1}) + d_{k0} + i_k + 1 \quad \text{for } k = 1, 2, \ldots, 4 \\ \phi_5(I_5) = 0 \end{cases} \tag{3.76c}$$

$$I_k = (i_k, i_{k+1}, \ldots, i_3, i_4) \tag{3.76d}$$

$$g_{k0}(i,j) = \frac{(-1)^i p_k(j)!}{i!(p_k(j)-i)!(d_{k,\,2-j} + i + 1)} \tag{3.76e}$$

$$g_{k1}(I_k) = \frac{(-1)^{i_k} \tilde{p}_k!}{i_k!(\tilde{p}_k - i_k)!\phi_k(I_k)} \tag{3.76f}$$

$$g_2(I_1) = \frac{(\phi_1(I_1) + d_{00})!\,\tilde{p}_0!}{(\phi_1(I_1) + d_{00} + \tilde{p}_0 + 1)!} \tag{3.76g}$$

$$f = \prod_{i=0}^{4} \frac{l_{i0}!\; l_{i1}!}{(l_{i0} + l_{i1} + 1)!}. \tag{3.76h}$$

As for θ_{LS}, we first define P_k as

$$P_k = \frac{l_{k0} + 1}{l_{k0} + l_{k1} + 2},$$

then use the rules shown in (3.18), (3.19), and (3.20). With $m = 5$ we have the definitions shown below.

$$g_{k0}^{(1)}(i,j) = \frac{d_{k,\,2-j} + i + 1}{d_{k,\,2-j} + i + 2}\, g_{k0}(i,j),$$
$$g_{k0}^{(2)}(i,j) = \frac{(-1)^i (p_k(j)+1)!}{i!(p_k(j)-i+1)!(d_{k,\,2-j} + i + 1)},$$

$$g_{k1}^{(1)}(I_k) = \frac{\phi_k(I_k)}{\phi_k(I_k)+1}\, g_{k1}(I_k),$$
$$g_{k1}^{(2)}(I_k) = \frac{(-1)^{i_k}(\tilde{p}_k+1)!}{i_k!(\tilde{p}_k - i_k + 1)!\phi_k(I_k)},$$

$$g_2^{(1)}(I_1) = \frac{\phi_1(I_1)+d_{00}+1}{\phi_1(I_1)+d_{00}+\tilde{p}_0+2}\, g_2(I_1) \quad \text{and}$$
$$g_2^{(2)}(I_1) = \frac{\tilde{p}_0+1}{\phi_1(I_1)+d_{00}+\tilde{p}_0+2}\, g_2(I_1)$$

for $k \in \{1,2,3,4\}$.

Now, using the definitions (3.16) and (3.21) for G and G_{st}, respectively, and from the equations (3.22), we get the formulas for the estimates of c_{st} and b_{s0} are as follows:

- Case $t = 0$:

$$\hat{c}_{00} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\cdot G \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\, g_{31}(I_3)\, g_{21}(I_2)\, g_{11}(I_1)\, g_2^{(1)}(I_1)$$
$$\hat{c}_{10} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\cdot G \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\, g_{31}(I_3)\, g_{21}(I_2)\, g_{11}^{(1)}(I_1)\, g_2^{(1)}(I_1)$$
$$\hat{c}_{20} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\cdot G \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\, g_{31}(I_3)\, g_{21}^{(1)}(I_2)\, g_{11}^{(1)}(I_1)\, g_2^{(1)}(I_1)$$
$$\hat{c}_{30} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\cdot G \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\, g_{31}^{(1)}(I_3)\, g_{21}^{(1)}(I_2)\, g_{11}^{(1)}(I_1)\, g_2^{(1)}(I_1)$$
$$\hat{c}_{40} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\cdot G \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}^{(1)}(I_4)\, g_{31}^{(1)}(I_3)\, g_{21}^{(1)}(I_2)\, g_{11}^{(1)}(I_1)\, g_2^{(1)}(I_1)$$

(3.77a)

where

$$G = \prod_{k=0}^{4}\left[\prod_{j=-1}^{1}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right] \tag{3.77b}$$

- Case $t \in \{1,2,3\}$:

$$\hat{c}_{0t} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\, G_{0t} \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\; g_{31}(I_3)\; g_{21}(I_2)\; g_{11}(I_1)\; g_2^{(2)}(I_1)$$

$$\hat{c}_{1t} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\, G_{1t} \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1+1} g_{41}(I_4)\; g_{31}(I_3)\; g_{21}(I_2)\; g_{11}^{(2)}(I_1)\; g_2(I_1)$$

$$\hat{c}_{2t} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\, G_{2t} \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2+1}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\; g_{31}(I_3)\; g_{21}^{(2)}(I_2)\; g_{11}(I_1)\; g_2(I_1)$$

$$\hat{c}_{3t} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\, G_{3t} \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3+1}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\; g_{31}^{(2)}(I_3)\; g_{21}(I_2)\; g_{11}(I_1)\; g_2(I_1)$$

$$\hat{c}_{4t} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f\, G_{4t} \sum_{i_4=0}^{\tilde{p}_4+1}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}^{(2)}(I_4)\; g_{31}(I_3)\; g_{21}(I_2)\; g_{11}(I_1)\; g_2(I_1)$$

(3.77c)

where

$$G_{st} = \left[\prod_{j=-1}^{1-t}\sum_{i=0}^{p_s(j)} g_{s0}(i,j)\right]\left[\sum_{i=0}^{p_s(2-t)} g_{s0}^{(1)}(i,2-t)\right]\left[\prod_{j=3-t}^{1}\sum_{i=0}^{p_s(j)+1} g_{s0}^{(2)}(i,j)\right] \cdot \prod_{\substack{k=0\\k\neq s}}^{4}\left[\prod_{j=-1}^{1}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right] \qquad \text{for } s\in\{0,1,2,3,4\}. \tag{3.77d}$$

As for the estimates $\hat{b}_{k0}$, for $k = 0, 1, \ldots, 4$, we have

$$\hat{b}_{k0} = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_k \cdot f \cdot G \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\; g_{31}(I_3)\; g_{21}(I_2)\; g_{11}(I_1)\; g_2(I_1),$$

(3.78)

Also, from the equation (3.4), we have

$$P\left(O^{1,n}\right) = \frac{1}{5} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_4=0}^{\tilde{p}_4}\sum_{i_3=0}^{\tilde{p}_3}\sum_{i_2=0}^{\tilde{p}_2}\sum_{i_1=0}^{\tilde{p}_1} g_{41}(I_4)\; g_{31}(I_3)\; g_{21}(I_2)\; g_{11}(I_1)\; g_2(I_1).$$

Here, as before, the product notation with its index going from larger value down to smaller value is defined to equal 1.

As for the exponents $K = \{k_{ij}\}$ (which can be expressed with the corresponding $\{d_{ij}\}$) and $L = \{l_{ik}\}$, $i, j \in \{0, 1, 2, 3, 4\}$ and $k \in \{0, 1\}$, given a state sequence $S^{1,n}$ and an

observation sequence $O^{1,n}$, let r_{ij} be as defined by the equation (3.27) for δ in Section 1. For $m_A = 5$, we have

$$R = \begin{pmatrix} r_{00} & r_{01} & r_{02} & r_{03} & r_{04} \\ r_{10} & r_{11} & r_{12} & r_{13} & r_{14} \\ r_{20} & r_{21} & r_{22} & r_{23} & r_{24} \\ r_{30} & r_{31} & r_{32} & r_{33} & r_{34} \\ r_{40} & r_{41} & r_{42} & r_{43} & r_{44} \end{pmatrix} = \frac{1}{24} \begin{pmatrix} 24 & -50 & 35 & -10 & 1 \\ 0 & 96 & -104 & 36 & -4 \\ 0 & -72 & 114 & -48 & 6 \\ 0 & 32 & -56 & 28 & -4 \\ 0 & -6 & 11 & -6 & 1 \end{pmatrix}. \tag{3.79}$$

Then, using the equations (3.41) in Section 1, we have the following: For $i, j \in \{1, 2, 3, 4\}$,

$$k_{00} = \sum_{i=1}^{4} \left[\sum_{j=1}^{4} k_{ij} - 2k_i + \left(\sum_{j=1}^{4} r_{ji} \right) (s_1^i + s_n^i) \right] + n - 1, \tag{3.80a}$$

$$k_{0j} = -\sum_{i=1}^{4} \left(k_{ij} + r_{ji} s_1^i \right) + k_j, \tag{3.80b}$$

$$k_{i0} = -\sum_{j=1}^{4} \left(k_{ij} + r_{ij} s_n^j \right) + k_i \tag{3.80c}$$

$$l_{00} = \sum_{i=1}^{4} (l_{i1} - k_i) - l_1 + n, \tag{3.80d}$$

$$l_{01} = -\sum_{i=1}^{4} l_{i1} + l_1, \quad \text{and} \tag{3.80e}$$

$$l_{i0} = -l_{i1} + k_i, \tag{3.80f}$$

where the values of r_{ij} are as in the equation (3.79). Also, with $m_A = 5$, we have

$$\omega = \left(K_4, K^{44}, L_4, s_1, s_n \right)$$

where

$$K_4 = (k_1,\, k_2,\, k_3,\, k_4)\,,$$

$$K^{44} = \begin{pmatrix} k_{11} & k_{12} & k_{13} & k_{14} \\ k_{21} & k_{22} & k_{23} & k_{24} \\ k_{31} & k_{32} & k_{33} & k_{34} \\ k_{41} & k_{42} & k_{43} & k_{44} \end{pmatrix}, \qquad \text{and}$$

$$L_4 = (l_{11},\, l_{21},\, l_{31},\, l_{41})\,. \tag{3.81}$$

Again, let $h_n(\omega)$ be the number of possible values of $s^{1,n} \in \Omega_n$ given a particular observation sequence $o^{1,n}$.

For the algorithm for finding h_n, we observe the following:

- If the next transition state s_{t+1} is 0, then
  - the last state s_n will be 0, and
  - the rest of the quantities will stay the same
- For $j \in \{1, 2, 3, 4\}$, if the next transition state s_{t+1} is j, then
  - the last state s_n will be j,
  - k_j (the number of j's in $s^{1,t+1}$) will be increased by 1,
  - k_{ij} (the number of ij's in $s^{1,t+1}$) will be increased by 1 if the current state s_t is i, and will stay the same otherwise, for $i \in \{1, 2, 3, 4\}$,
  - l_{j1} (the number of times 1 is observed when the state is j, given $s^{1,t+1}$ and $o^{1,t+1}$) will be increased by 1 if $o_{t+1} = 1$, and stays the same otherwise, and
  - the rest of the quantities will stay the same.

The algorithm is as shown below. Using the algorithm below, we first find the values of $h_n(\omega)$ for ω-values that corresponds to $s_n \in \left\{\overline{\Omega}_n\right\}$ that starts with $s_1 = 0$. Then, for each

value i of states $\{1, 2, 3, 4\}$, we interchange the states i and 0 to evaluate the integration necessary for the LSE. (For example, if $k_{30} = x$ results from the original ω then let $k_{3i} = x$, and so on. See the description in Section 1.1.)

Algorithm for finding h_n

Let $h_1(0_4, 0^{44}, 0_4, 0, 0) = 1$, where 0_4 is a length-4 zero vector and 0^{44} is a 4×4 zero matrix.

for t from 1 to $n-1$

with all $\omega = (K_4, K^{44}, L_4, 0, s_t)$ such that $h_t(\omega) > 0$

Increment $h_{t+1}(K_4, K^{44}, L_4, 0, 0)$ by the value of $h_t(\omega)$ (for the case $s_{t+1} = 0$).

for s_{t+1} from 1 to 4

Obtain $\hat{\omega}$ from ω by incrementing

(*i*) $k_{s_{t+1}}$ in K_4 by one,

(*ii*) $k_{s_t, s_{t+1}}$ in K^{44} by one, and

(*iii*) $l_{s_{t+1}, 1}$ in L_4 by o_{t+1},

then replacing s_t by s_{t+1}.

Increment $h_{t+1}(\hat{\omega})$ by the value of $h_t(\omega)$.

end for

end for

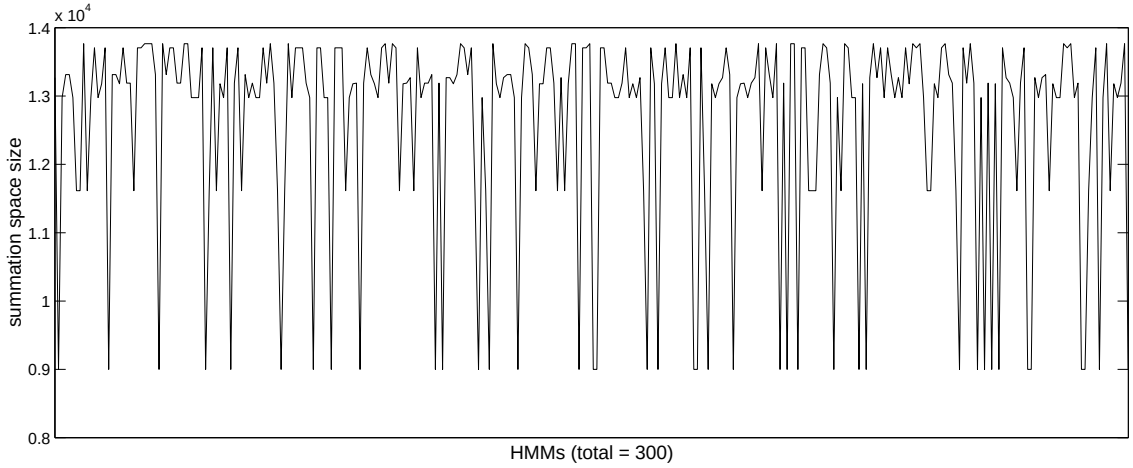


Figure 2. Change in the size of $\overline{\Omega}_7$ with $m_A = 5$ and sequence length 6, over 150 randomly generated observation sequences.

Figure 2 shows the experimental results for the size of $\overline{\Omega}_n$ for $n = 7$. As done in the case of the state space size $m_A = 2$, three hundred $\theta = (A, B, \pi)$ values are randomly selected, but π is fixed as $\frac{1}{5}$. Then, for each θ, a state sequence and an observation sequence of length 5 are generated. On average, the size of $\overline{\Omega}_7$ was 12600.2, which is smaller than $5^7 = 78125$, but not as significantly smaller as in the case of $m_A = 2$. The size reduction gets more significant as the sequence length increases.

4.2. Covariance Matrix. In order to find the covariance matrix, we need to find $E\left(c_{s_1 t_1}\, c_{s_2 t_2} \mid O^{1,n}\right)$, $E\left(b_{s_1 0}\, b_{s_2 0} \mid O^{1,n}\right)$, and $E\left(c_{s_1 t}\, b_{s_2 0} \mid O^{1,n}\right)$.

Using the definitions (3.47) and (3.48) in Section 1.2, we first define functions shown below.

$$g_{k0}^{(11)}(i,j) = \frac{d_{k,\,2-j} + i + 1}{d_{k,\,2-j} + i + 3}\, g_{k0}(i,j), \tag{3.82a}$$

$$g_{k0}^{(22)}(i,j) = \frac{(-1)^i (p_k(j)+2)!}{i!(p_k(j)-i+2)!(d_{k,\,2-j}+i+1)}, \tag{3.82b}$$

$$g_{k1}^{(11)}(I_k) = \frac{\phi_k(I_k)}{\phi_k(I_k)+2}\, g_{k1}(I_k), \tag{3.82c}$$

$$g_{k1}^{(22)}(I_k) = \frac{(-1)^{i_k}(p_k(2)+2)!}{i_k!(p_k(2)-i_k+2)!\phi_k(I_k)}, \tag{3.82d}$$

$$g_{k1}^{(12)}(I_k) = \frac{(-1)^{i_k}(p_k(2)+1)!}{i_k!(p_k(2)-i_k+1)!\,(\phi_k(I_k)+1)}, \tag{3.82e}$$

$$g_2^{(11)}(I_1) = \frac{\phi_1(I_1)+d_{00}+2}{\phi_1(I_1)+d_{00}+p_0(2)+3}\, g_2^{(1)}(I_1), \tag{3.82f}$$

$$g_2^{(22)}(I_1) = \frac{p_{02}+2}{\phi_1(I_1)+d_{00}+p_0(2)+3}\, g_2^{(2)}(I_1), \quad \text{and} \tag{3.82g}$$

$$g_2^{(12)}(I_1) = \frac{p_0(2)+1}{\phi_1(I_1)+d_{00}+p_0(2)+3}\, g_2^{(1)}(I_1). \tag{3.82h}$$

Also, define $G_{s_1t_1,\,s_2t_2}$ as follows: If $s_1 = s_2 = s$ and $t_1 = t_2 = t$, then

$$G_{st,\,st} = \left[\prod_{j=-1}^{1-t}\sum_{i=0}^{p_s(j)} g_{s0}(i,j)\right]\left[\sum_{i=0}^{p_s(2-t)} g_{s0}^{(11)}(i,2-t)\right]\left[\prod_{j=3-t}^{1}\sum_{i=0}^{p_s(j)+1} g_{s0}^{(22)}(i,j)\right] \cdot \prod_{\substack{k=0\\k\neq s}}^{4}\left[\prod_{j=-1}^{1}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]. \tag{3.83a}$$

If $s_1 = s_2 = s$ and $t_1 \neq t_2$, let $T = max\{t_1,\, t_2\}$ and $t = min\{t_1,\, t_2\}$. Then

$$G_{st_1,\,st_2} = \left[\prod_{j=-1}^{1-T}\sum_{i=0}^{p_s(j)} g_{s0}(i,j)\right]\left[\sum_{i=0}^{p_s(2-T)} g_{s0}^{(1)}(i,2-T)\right]\left[\prod_{j=3-T}^{1-t}\sum_{i=0}^{p_s(j)+1} g_{s0}^{(2)}(i,j)\right] \cdot \left[\sum_{i=0}^{p_s(2-t)} g_{s0}^{(12)}(i,2-t)\right]\left[\prod_{j=3-t}^{1}\sum_{i=0}^{p_s(j)+1} g_{s0}^{(22)}(i,j)\right] \cdot \prod_{\substack{k=0\\k\neq s}}^{4}\left[\prod_{j=-1}^{1}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]. \tag{3.83b}$$

If $s_1 \neq s_2$ then

$$G_{s_1t_1,\,s_2t_2} = \left[\prod_{j=-1}^{1-t_1}\sum_{i=0}^{p_{s_1}(j)} g_{s_1,\,0}(i,j)\right]\left[\sum_{i=0}^{p_{s_1}(2-t_1)} g^{(1)}_{s_1,\,0}(i,2-t_1)\right]$$
$$\cdot\left[\prod_{j=3-t_1}^{1}\sum_{i=0}^{p_{s_1}(j)+1} g^{(2)}_{s_1,\,0}(i,j)\right]\left[\prod_{j=-1}^{1-t_2}\sum_{i=0}^{p_{s_2}(j)} g_{s_2,\,0}(i,j)\right]$$
$$\cdot\left[\sum_{i=0}^{p_{s_2}(2-t_2)} g^{(1)}_{s_2,\,0}(i,2-t_2)\right]\left[\prod_{j=3-t_2}^{1}\sum_{i=0}^{p_{s_2}(j)+1} g^{(2)}_{s_2,\,0}(i,j)\right]$$
$$\cdot\prod_{\substack{k=0\\ k\neq s_1,\,s_2}}^{4}\left[\prod_{j=-1}^{1}\sum_{i=0}^{p_k(j)} g_{k0}(i,j)\right]. \tag{3.83c}$$

For $s_1, s_2 \in \{1,2,\ldots,4\}$ and $t_1, t_2 \in \{1,2,\ldots,3\}$, we have the expected values as shown below. Note that, in order to avoid listing all the possible combinations of values of s_1, s_2, t_1, t_2, same quantity is sometimes written more than once or some subscriptions are not in the range.

For example, in order to find $E\left(c_{10}c_{20} \mid O^{1,n}\right)$, if we simply plug in $s_1 = 1$ and $s_2 = 2$ in the equation (3.84g) below, then we get

$$E\left(c_{10}c_{20} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)}\sum_{s^{1,n}\in\Omega_n} f\cdot G\sum_{i_4=0}^{p_4(2)}\sum_{i_3=0}^{p_3(2)}\sum_{i_3=0}^{p_2(2)}\sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31}$$
$$\cdots g_{31}\; g^{(1)}_{21}\; g^{(1)}_{11}\cdots g^{(1)}_{21}\; g^{(11)}_{11}\; g^{(11)}_{01}\; g^{(11)}_{11}\cdot g^{(11)}_{2},$$

since $Q = max\{1,2\} = 2$ and $q = min\{1,2\} = 1$. So, as it is, it has $g^{(11)}_{01}$, which subscript (i.e., 01) is out of range, and several expressions that appear twice. But, if we (i) ignore the expressions with invalid subscript and (ii), in case of the same expression appearing more than once, ignore the factors in between, then we get

$$E\left(c_{10}c_{20} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)}\sum_{s^{1,n}\in\Omega_n} f\cdot G\sum_{i_4=0}^{p_4(2)}\sum_{i_3=0}^{p_3(2)}\sum_{i_3=0}^{p_2(2)}\sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31}\; g^{(1)}_{21}\; g^{(11)}_{11}\cdot g^{(11)}_{2}.$$

Now the formulas are as shown below.

$$E\left(c_{00}\, c_{00} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{11} \cdot g_2^{(11)}. \tag{3.84a}$$

$$E\left(c_{00}\, c_{0t_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{11} \cdot g_2^{(12)}. \tag{3.84b}$$

$$E\left(c_{00}\, c_{s_20} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{s_2+1,\,1}\; g_{s_2,\,1}^{(1)}\; g_{s_2-1,\,1}^{(1)} \cdots g_{11}^{(1)} \cdot g_2^{(11)}. \tag{3.84c}$$

$$E\left(c_{00}\, c_{s_2t_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_2\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{s_2+1}=0}^{p_{s_2+1}(2)} \sum_{i_{s_2}=0}^{p_{s_2}(2)+1} \sum_{i_{s_2-1}=0}^{p_{s_2-1}(2)} \cdots \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31}\; \cdots g_{s_2+1,\,1}\; g_{s_2,\,1}^{(2)}\; g_{s_2-1,\,1} \cdots g_{11} \cdot g_2^{(1)} \tag{3.84d}$$

$$E\left(c_{0t_1}\, c_{0t_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0\,t_1,\,0\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{11} \cdot g_2^{(22)}. \tag{3.84e}$$

$$E\left(c_{0t_2}\, c_{s_20} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{s_2+1,\,1}\; g_{s_2,\,1}^{(1)}\; g_{s_2-1,\,1}^{(1)} g_{11}^{(1)} \cdot g_2^{(12)}. \tag{3.84f}$$

$$E\left(c_{0t_1}\, c_{s_2t_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{0\,t_1,\,s_2\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{s_2+1}=0}^{p_{s_2+1}(2)} \sum_{i_{s_2}=0}^{p_{s_2}(2)+1} \sum_{i_{s_2-1}=0}^{p_{s_2-1}(2)}$$
$$\cdots \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31}\; \cdots g_{s_2+1,\,1}\; g^{(2)}_{s_2,\,1}\; g_{s_2-1,\,1} \cdots g_{11} \cdot g^{(2)}_2$$

Let $Q = max\{s_1, s_2\}$ and $q = min\{s_1, s_2\}$ then

$$E\left(c_{s_10}\, c_{s_20} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_3=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31}$$
$$\cdots g_{Q+1,\,1}\; g^{(1)}_{Q1}\; g^{(1)}_{Q-1,\,1} \cdots g^{(1)}_{q+1,\,1}\; g^{(11)}_{q1}\; g^{(11)}_{q-1,\,1} \cdots g^{(11)}_{11} \cdot g^{(11)}_2. \tag{3.84g}$$

If $s_1 = s_2 = s$ then

$$E\left(c_{s0}\, c_{st_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{s+1}=0}^{p_{s+1}(2)} \sum_{i_s=0}^{p_s(2)+1} \sum_{i_{s-1}=0}^{p_{s-1}(2)}$$
$$\cdots \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{s+1,\,1}\; g^{(12)}_{s1}\; g^{(1)}_{s-1,\,1} \cdots g^{(1)}_{11} \cdot g^{(1)}_2. \tag{3.84h}$$

If $s_1 > s_2$ then

$$E\left(c_{s_10}\, c_{s_2t_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_2\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{s_2+1}=0}^{p_{s_2+1}(2)} \sum_{i_{s_2}=0}^{p_{s_2}(2)+1} \sum_{i_{s_2-1}=0}^{p_{s_2-1}(2)}$$
$$\cdots \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{s_1+1,\,1}\; g^{(1)}_{s_1,\,1}\; g^{(1)}_{s_1-1,\,1}$$
$$\cdots g^{(1)}_{s_2+1,\,1}\; g^{(12)}_{s_2,\,1}\; g^{(1)}_{s_2-1,\,1} \cdots g^{(1)}_{11} \cdot g^{(1)}_2. \tag{3.84i}$$

If $s_1 < s_2$ then

$$E\left(c_{s_10}\, c_{s_2t_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_2\,t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{s_2+1}=0}^{p_{s_2+1}(2)} \sum_{i_{s_2}=0}^{p_{s_2}(2)+1} \sum_{i_{s_2-1}=0}^{p_{s_2-1}(2)}$$
$$\cdots \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31}\; \cdots g_{s_2+1,\,1}\; g^{(2)}_{s_2,\,1}\; g_{s_2-1,\,1}$$
$$\cdots g_{s_1+1,\,1}\; g^{(1)}_{s_1,\,1}\; g^{(1)}_{s_1-1,\,1} \cdots g^{(1)}_{11} \cdot g^{(1)}_2. \tag{3.84j}$$

If $s_1 = s_2 = s$ then

$$E\left(c_{st_1}\, c_{st_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{st_1,\, st_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{s+1}=0}^{p_{s+1}(2)} \sum_{i_s=0}^{p_s(2)+2} \sum_{i_{s-1}=0}^{p_{s-1}(2)} \cdots \sum_{i_1=0}^{p_1(2)} g_{41}\, g_{31} \cdots g_{s+1,\,1}\, g_{s1}^{(22)}\, g_{s-1,\,1} \cdots g_{11} \cdot g_2. \tag{3.84k}$$

If $s_1 \neq s_2$ then, with $Q = max\{s_1, s_2\}$ and $q = min\{s_1, s_2\}$,

$$E\left(c_{s_1t_1}\, c_{s_2t_2} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} f \cdot G_{s_1t_1,\, s_2t_2} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{Q+1}=0}^{p_{Q+1}(2)} \sum_{i_Q=0}^{p_{Q2}+1} \sum_{i_{Q-1}=0}^{p_{Q-1}(2)} \cdots \sum_{i_{q+1}=0}^{p_{q+1}(2)} \sum_{i_q=0}^{p_q(2)+1} \sum_{i_{q-1}=0}^{p_{q-1}(2)} \cdots \sum_{i_1=0}^{p_1(2)} g_{41}\, g_{31}\, \cdots g_{Q+1,\,1}\, g_{Q1}^{(2)}\, g_{Q-1,\,1} \cdots g_{q+1,\,1}\, g_{q1}^{(2)}\, g_{q-1,\,1} \cdots g_{11} \cdot g_2$$

Similarly, as for the expected values $E\left(b_{s_1,\,0} b_{s_2,\,0} \mid O^{1,n}\right)$, $s_1, s_2 \in \{0, 1, \ldots, 4\}$, we define $\widetilde{P}_s$ in the same way as defined in (3.50).

$$\widetilde{P}_s = \frac{l_{s0} + 2}{l_{s0} + l_{s1} + 3} \tag{3.85}$$

Then, if $s_1 = s_2 = s$, we have

$$E\left(b_{s0}\, b_{s0} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} \widetilde{P}_s P_s \cdot f \cdot G \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\, g_{31} \cdots g_{11} \cdot g_2, \tag{3.86a}$$

and if $s_1 \neq s_2$, then

$$E\left(b_{s_1,\,0} b_{s_2,\,0} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_{s_1} P_{s_2} \cdot f \cdot G \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\, g_{31} \cdots g_{11} \cdot g_2. \tag{3.86b}$$

As for the expected values in the form $E\left(c_{s_1 t}\, b_{s_2 0} \mid O^{1,n}\right)$, we have the following:

$$E\left(c_{s_1 0}\, b_{s_2 0} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_{s_2}\, f \cdot G \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{s_1+1,\,1}\; g^{(1)}_{s_1,\,1}\; g^{(1)}_{s_1-1,\,1} \cdots g^{(1)}_{11} \cdot g^{(1)}_{2}. \tag{3.87a}$$

$$E\left(c_{00}\, b_{s_2 0} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_{s_2}\, f \cdot G \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{11} \cdot g^{(1)}_{2}. \tag{3.87b}$$

$$E\left(c_{s_1 t}\, b_{s_2 0} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_{s_2} f \cdot G_{s_1\, t} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \cdots \sum_{i_{s_1+1}=0}^{p_{s_1+1}(2)} \sum_{i_{s_1}=0}^{p_{s_1}(2)+1} \sum_{i_{s_1-1}=0}^{p_{s_1-1}(2)} \cdots \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31}\; \cdots g_{s_1+1,\,1}\; g^{(2)}_{s_1,\,1}\; g_{s_1-1,\,1} \cdots g_{11} \cdot g_2. \tag{3.87c}$$

$$E\left(c_{0t}\, b_{s_2 0} \mid O^{1,n}\right) = \frac{1}{5P\left(O^{1,n}\right)} \sum_{s^{1,n}\in\Omega_n} P_{s_2}\, f \cdot G_{0t} \sum_{i_4=0}^{p_4(2)} \sum_{i_3=0}^{p_3(2)} \sum_{i_2=0}^{p_2(2)} \sum_{i_1=0}^{p_1(2)} g_{41}\; g_{31} \cdots g_{11} \cdot g^{(2)}_{2}. \tag{3.87d}$$

5. LSE with Particle Filter

Particle filters are sequential Monte Carlo methods, which can be applied to any probabilistic state-space model, also known as sequential importance sampling algorithms [2, 10]. Considered here is one of the basic implementation of the particle filter method. This is a method that is widely used in various stochastic systems, although most often it is used for prediction not for parameter estimation.

For parameter estimation, we consider a vector of state S_t and parameter set θ, which we will call as $p_t = (S_t, \theta)$, and its assigned normalized weight $\bar{w}_t$. The idea is that if

(1) we sample from p_t to get $\left\{p_t^{(i)}\right\}_{i=1}^{N}$, (2) assign a normalized weight $\bar{w}_t^{(i)}$ to each $p_t^{(i)}$ so that, given an observation sequence $O^{1,n}$, the more likely $p_t^{(i)}$ is the heavier the weight is, and then (3) resample from $\left\{p_t^{(i)}\right\}_{i=1}^{N}$ so that a sample with heavier weight is more likely to be sampled again; then we will eventually find an approximation of $P\left(p_t \mid O^{1,n}\right)$. The samples in $\left\{p_t^{(i)}\right\}_{i=1}^{N}$ are called particles.

Given an observation sequence $O^{1,n}$, the expected value of p_t can be obtained by

$$E\left(p_t\right) = \int p_t \, P\left(p_t \mid O^{1,n}\right) dp_t = \sum_{i=1}^{N} \bar{w}_t^{(i)} p_t^{(i)}.$$

5.1. Algorithm of the Particle Filter. Let $\theta = \{\pi, A, B\}$ be the parameter set, $\bar{w}_t^{(i)}$ be the normalized weight for the i-th particle $p_t^{(i)} = \left(S_t^{(i)}, \theta_t^{(i)}\right)$ for discrete time t, and N is the total number of particles. Assume a particular observation sequence, $o^{1,n}$, is given. In below, the outline is given first, then the detailed description is given as Algorithms 1, 2, and 3.

MAIN (Particle Filter)

(1) Randomly pick the first set of particles $\left\{p_1^{(i)}\right\}_{i=1}^{N} = \left\{\left(s_1^{(i)}, \theta_1^{(i)}\right)\right\}_{i=1}^{N}$, then attach weights $\bar{w}_1^{(i)}$ that are computed using $s_1^{(i)}$, $\theta_1^{(i)}$ and o_1. [Algorithm 1]

(2) Resample from $\left\{p_1^{(i)}\right\}_{i=1}^{N}$ to have a new set of particles $\left\{\hat{p}_1^{(i_k)}\right\}_{i_k=1}^{N}$, according to the assigned weights $\bar{w}_1^{(i)}$. [Algorithm 2]

(3) Repeat the procedures below for each o_t, t from 2 to n

Pick the next set $\left\{p_t^{(i)}\right\}_{i=1}^{N}$ according to the values of current set $\left\{\hat{p}_{t-1}^{(i_k)}\right\}_{i_k=1}^{N}$ and o_t, while determining each weight $\bar{w}_t^{(i)}$ using $s_t^{(i)}$, $\theta_t^{(i)}$ and o_t. [Algorithm 3]

Resample from $\left\{p_t^{(i)}\right\}_{i=1}^{N}$ to have a new set of particles $\left\{\hat{p}_t^{(i_k)}\right\}_{i_k=1}^{N}$, according to the attached weights $\bar{w}_t^{(i)}$. [Algorithm 2]

(4) Let the average value of the parameters in $\left\{\theta_n^{(i)}\right\}_{i=1}^{N}$ be $\hat{\theta}$, which will approximate $\theta_{LS} = E(\theta \mid O^{1,n})$.

Algorithm 1: Pick the first set of samples.

Repeat until N particles $\left\{p_1^{(i)}\right\}_{i=1}^{N} = \left\{\left(s_1^{(i)}, \theta_1^{(i)}\right)\right\}_{i=1}^{N}$ are picked:

Pick $\theta_1^{(i)} = \left\{\pi_1^{(i)}, A_1^{(i)}, B_1^{(i)}\right\}$ assuming the uniform distribution.

Pick $s_1^{(i)} \sim \pi_1^{(i)}(s_1^{(i)})$, where $\pi_1^{(i)}(n)$ corresponds to $P(S_1 = n)$.

Let $w_1^{(i)} = B_1^{(i)}(s_1^{(i)}, o_1)$, where $B_1^{(i)}(n, m)$ corresponds to $P(O_t = m \mid S_t = n)$.

Normalize $w_1^{(i)}$ to get $\bar{w}_t^{(i)}$

Algorithm 2: Resample.

Repeat until N particles $\left\{\hat{p}_t^{(i_k)}\right\}_{i_k=1}^{N}$ are picked:

Pick $p_t^{(i)} = \left\{\left(s_t^{(i)}, \theta_t^{(i)}\right)\right\} \sim \bar{w}_t^{(i)}$ from $\left\{p_t^{(i)}\right\}_{i=1}^{N}$, *with replacement.*

Denote this $p_t^{(i)}$ as $\hat{p}_t^{(i_k)} = \left(\hat{s}_t^{(i_k)}, \hat{\theta}_t^{(i_k)}\right)$. Let $\hat{w}_t^{(i_k)} = \frac{1}{N}$.

Algorithm 3: Find the next set of samples.

Repeat until N particles $\left\{p_t^{(i)}\right\}_{i=1}^{N}$ are picked:

Let $\theta_t^{(i)} = \left\{\pi_t^{(i)}, A_t^{(i)}, B_t^{(i)}\right\} = \hat{\theta}_{t-1}^{(i_k)} + \epsilon$, where ϵ is any *small* noise.

Pick $s_t^{(i)} \sim A_t^{(i)}\left(\hat{s}_{t-1}^{(i_k)}, s_t^{(i)}\right)$, where $A_t(m, n)$ corresponds to $P(S_t = n \mid S_{t-1} = m)$.

Let $p_t^{(i)} = \left(s_t^{(i)}, \theta_t^{(i)}\right)$.

Let $w_t^{(i)} = \hat{w}_{t-1}^{(i_k)} B_t^{(i)}\left(s_t^{(i)}, o_t\right)$.

Normalize $w_t^{(i)}$ to get $\bar{w}_t^{(i)}$.

Note that the noise used in Algorithm 3 is used to make the particles to spread over the space.

5.2. Results. With state space size $m_A = 2$, 300 parameter sets θ_i, $i = 1, 2, \ldots, 300$ are randomly chosen, and, using each θ_i, one set of state and observation sequences of length 20 are generated. Then the particle filter estimations are found using 500, 1000, 5000, and 10000 particles, and the Euclidean distance between the least square estimate $\hat{\theta}_{LS,i}$ and the particle filter estimates have been computed. Figure 3 shows a part of the results of this experiment. The Euclidean distances between the exact LSE and its particle filter approximation are plotted for $\theta_1, \theta_2, \ldots, \theta_{100}$ and for the two types of particle filter estimates, one using 500 and the other using 10000 particles.

The average distances found are approximately 0.424, 0.322, 0.240, and 0.250, while the average variances are 0.216, 0.531, 0.886, and 1.241, for the cases 500, 1000, 5000, and 10000 particles used, respectively. On average, as expected, the distance tends to go down as the number of particles increases. To get the idea of how close the LSE and its particle filter approximation are; for example, if the exact LSE is $\hat{\theta}_{LS} = (a, b, x, y, r) = (0.64, 0.30, 0.33, 0.63, 0.60)$ and the particle filter approximation of LSE is $\hat{\theta}_{PF} = (0.62, 0.32, 0.47, 0.51, 0.44)$ then the distance is approximately 0.246.

Similar experiment is implemented with the state space size $m_A = 5$. The average distances between the particle filter estimation and LSE estimation found are approximately 0.626, 5.725, and 5.202, while the average variances are 1.116, 1.182, and 1.208 for the cases 500, 1000, and 5000 particles used, respectively. Again, as the number of particles increased, the distance decreased as in the case of m_A above; however, the variance stayed almost the same. See Figure 4 for the distances with the first 100 θ values and for the particle number 500 and 5000.

Further experimental results about the particle filter method, in comparison with the B-W and LSE estimations, are shown in Chapter 4 below.

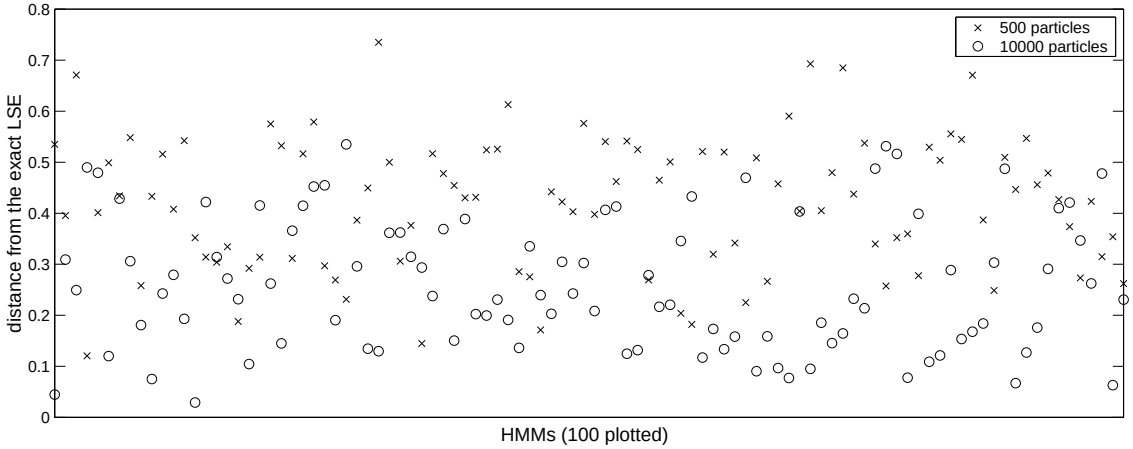


Figure 3. The distance between the exact LSE estimation and the particle filter approximation of LSE for the cases 500 particles are used (cross) and 1000 particles are used (circle) when $m_A = 2$.

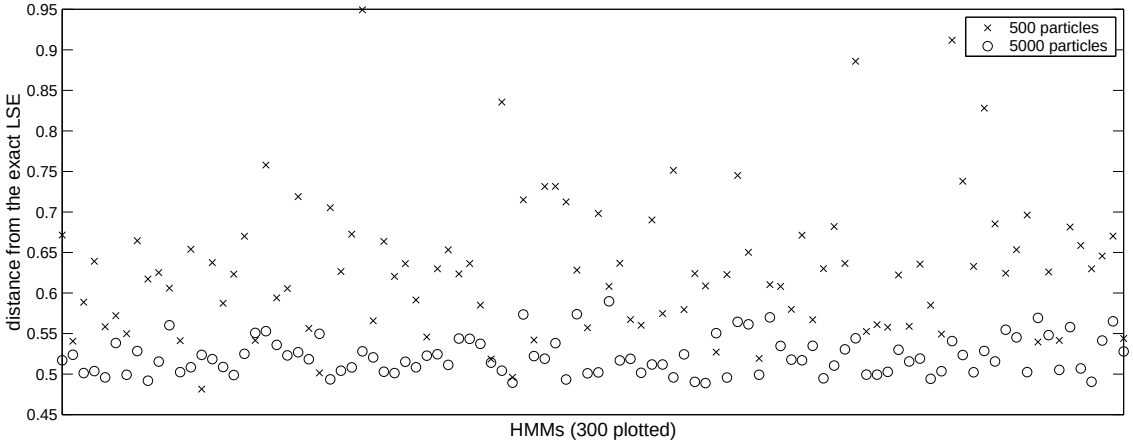


Figure 4. The distance between the exact LSE estimation and the particle filter approximation of LSE for the cases 500 particles are used (cross) and 5000 particles are used (circle) when $m_A = 5$.

CHAPTER 4

Comparing the Maximum Likelihood Baum-Welch Algorithm with Least Square Estimation

Shown in this chapter are some of the simulation results that compares the B-W, LSE, and the particle filter estimation.

1. Does Baum-Welch Algorithm Maximize the Likelihood?

The answer to the above question depends on the observation sequence the likelihood is about; i.e., whether it is the likelihood of the observation sequence that is used for the parameter estimation or of any observation sequence that is generated by the actual HMM.

1.1. Yes, the Baum-Welch Algorithm maximizes the likelihood. As a result of experiments described below, we found the B-W algorithm does maximizes the likelihood for a given observation sequence the majority of the time.

1.1.1. *Correlation Coefficients.* Five hundred values of θ_i, $i = 1, 2, \ldots, 500$ are randomly picked and one set of state and observation sequences of length 30 are generated for each i, which we denote as $o_i^{1,30}$. Then, 40 B-W estimates $\hat{\theta}_i(j)$, $j = 1, 2, \ldots, 40$, of θ_i are found using randomly picked initial estimates for each i. After grouping the initial

estimates that converged to a same estimate, say $\hat{\theta}_i(k)$ for some $k \in \{1, 2, \ldots, 40\}$, the correlation coefficient of the likelihood, $P\left(o_i^{1,30} \mid \hat{\theta}_i(k)\right)$, and the number of initial estimates that converges to $\hat{\theta}_i(k)$ are found. This value would approximate the correlation coefficient of the fixed point $\hat{\theta}_i(k)$ and the size of its basin. See Figure 5 for the results. Although the overall average correlation coefficient is approximately 0.55 over the estimates for 500 observation sequences $o_i^{1,30}$ that are generated using θ_i, you can see the values are positive and high most of the time. In fact, 33.6 percent of the time the correlation coefficient is higher than 0.9.

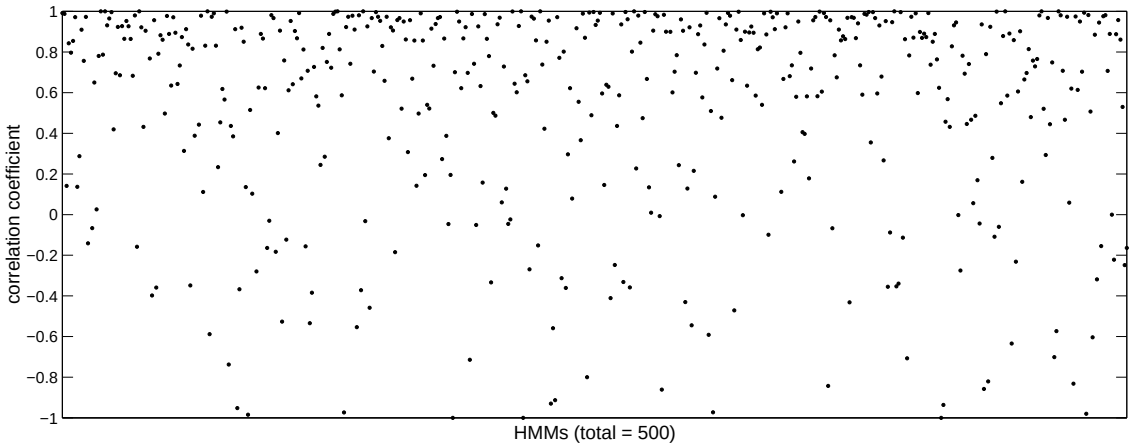


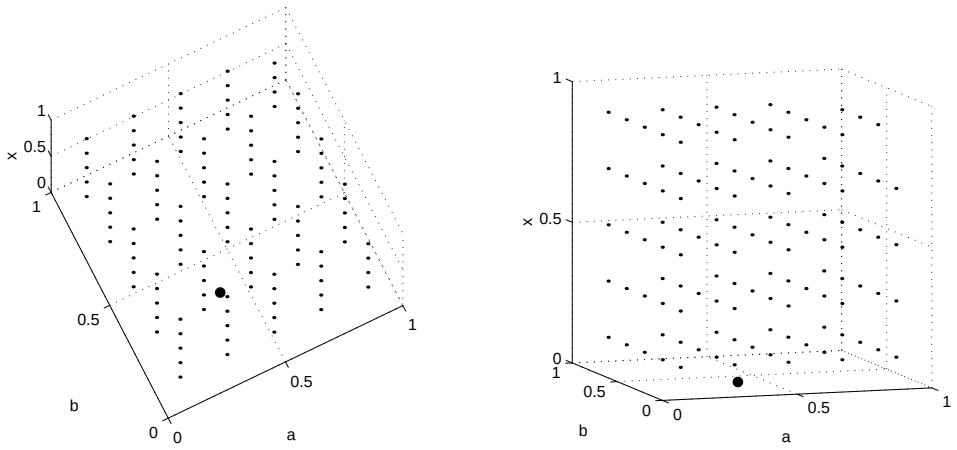
Figure 5. Correlation coefficients of $P\left(o^{1,30} \mid \hat{\theta}\right)$ and the number of initial values that converge to $\hat{\theta}$

A typical case in which the basin for the fixed point with the highest likelihood is dominant is shown in Figures 6 and 7. A parameter set $\theta = (a, b, x, y, r) = (0.64, 0.36, 0.58, 0.87, 0.61)$ is used to generate length-30 transition and observation sequences. Out of $5^5 = 3125$ initial parameter values used, which are picked with equal distance to each other for the purpose of plotting, 1497 converged to a fixed point $\theta_0 \approx (0.40, 0.36, 0.00, 0.00, 1.00)$, while other 1497 converged to a fixed point $\tilde{\theta}_0 \approx$

$(0.36, 0.40, 1.00, 1.00, 0.00)$, which is in symmetry to θ_0.

The fixed point θ_0 is plotted as a large filled-in circle, and its corresponding initial parameter points are plotted as a small dot. Since the y- and r-values are chosen to be close to this θ_0, you can see the initial points that converge to this fixed point are very dominant in the view in Figure 6. If we move the ranges for y and r away from θ_0 by 0.2, then the initial points that converge to $\tilde{\theta}_0$, plotted with a start mark, start to appear on the edge in Figure 7. Also, it is typical that the fixed point has some coordinates 0 or 1.

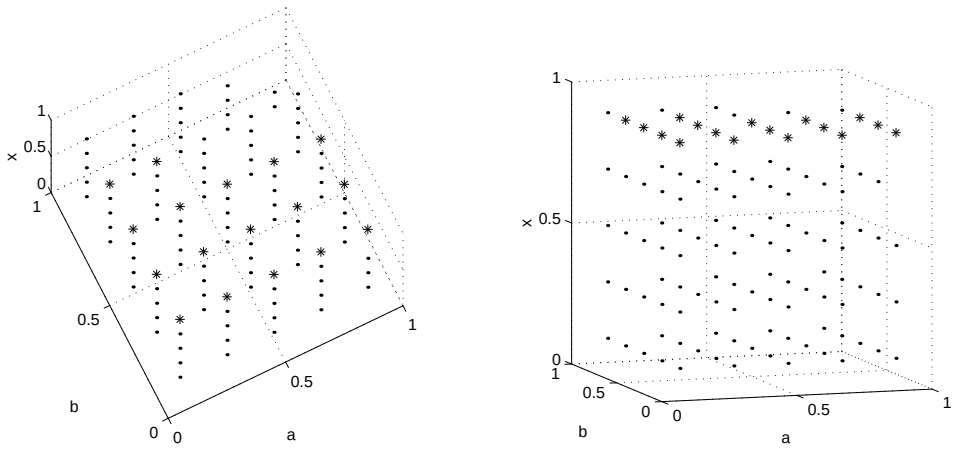
The likelihood $P\left(O^{1,n} \mid \theta_0\right) = P\left(O^{1,n} \mid \tilde{\theta}_0\right) \approx 4.5\ 10^{-9}$ is the highest compared to the one with other fixed points, and so the B-W has maximized the likelihood for this case.



(a) top view (b) side view

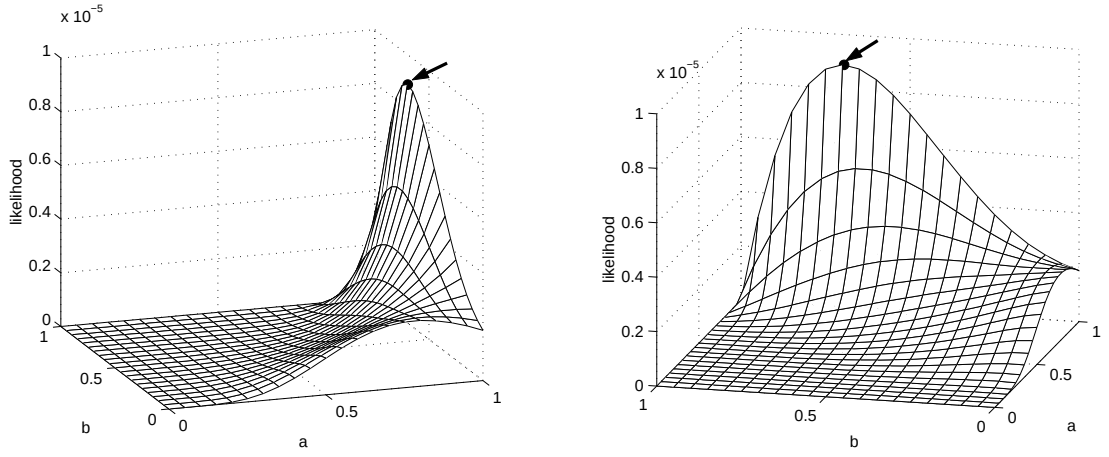
Figure 6. A fixed point (a big dot) and the initial values that converge to the fixed point (small dots) with $0 < y < 0.2$ and $0.8 < r < 1$.

Incidentally, the LSE for the same observation sequence is $\hat{\theta}_{LS}$ = (0.67, 0.30, 0.48, 0.55, 0, 50) and the corresponding likelihood $P\left(O^{1,n} \mid \hat{\theta}_{LS}\right)$ is approximately $1.0\ 10^{-9}$, while with the true parameter θ, we have $P\left(O^{1,n} \mid \theta\right) \approx 0.9\ 10^{-9}$.



(a) top view (b) side view

Figure 7. The initial values that converge to a fixed point shown in Fig. 6 (small dots), and the initial values that converge to another fixed point in symmetry (stars) with $0.2 < y < 0.4$ and $0.8 < r < 1$.



(a) a view from a-axis (b) a view from b-axis

Figure 8. The distribution of likelihood $P\left(o^{1,20} \mid \hat{\theta}_{BW}\right)$ and a fixed point $\hat{\theta}_{BW}$ (pointed by an arrow), where x-, y-, and r-values are fixed as equal to those of the fixed point.

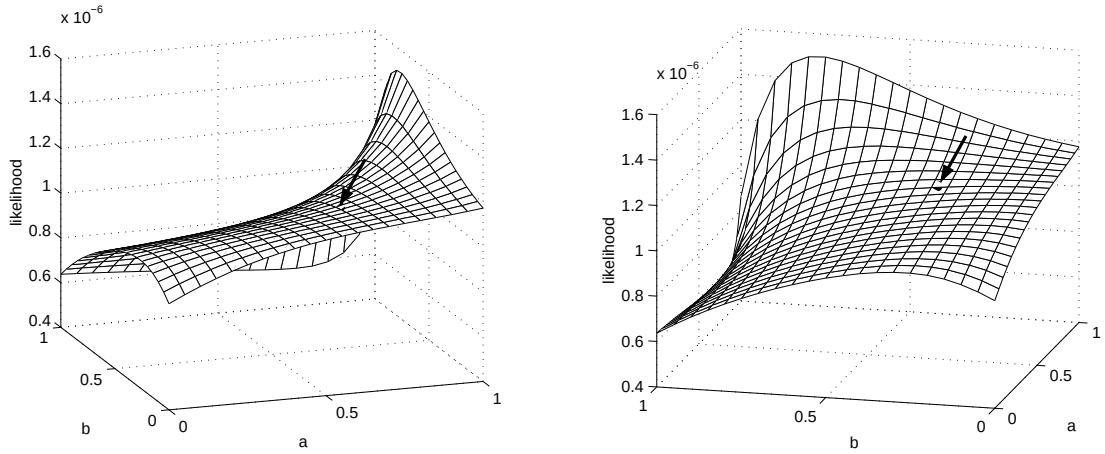
1.1.2. *Distribution of Likelihood.* A typical distribution of the likelihood near a fixed point is shown in Figure 8. An observation sequence

$$o^{1,20} = (1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0)$$

is generated using the parameter set $\theta = (a, b, x, y, r) = (0.86, 0.23, 0.45, 0.57, 0.52)$, and the B-W algorithm found two fixed points with the largest basin, which are $\hat{\theta}_{BW} = (a, b, x, y, r)$ $\approx (1.00, 0.70, 0.66, 1.00, 0.00)$ and its corresponding point in symmetry. The size and shape of the basin for this fixed point is very similar to the one shown in Figures 6 and 7, though the HMM used for each experiment (i.e., θ-values) are different. We see that, at least if x-, y-, and r-values are fixed as those of $\hat{\theta}_{BW}$, it found the absolute maximum $P\left(o^{1,20} \mid \hat{\theta}_{BW}\right)$ $\approx 8.8910^{-6}$. Furthermore, we see the fixed point is on the border, which is very usual for the B-W estimators, and which results in having almost-zero likelihood values for other observation sequences that are generated by the same θ, quite often. This issue of almost-zero probability is discussed later.

The LSE for the above observation sequence $o^{1,20}$ is $\hat{\theta}_{LS} = (a, b, x, y, r) \approx$ $(0.66, 0.33, 0.58, 0.55, 0.48)$, and $P\left(o^{1,20} \mid \hat{\theta}_{LS}\right) \approx 1.1010^{-6}$. Figure 9 shows the likelihood distribution near this estimate.

1.2. No, the Baum-Welch Algorithm does not maximize the likelihood. In the preceding section, we found that the B-W algorithm does maximizes the likelihood of a given observation sequence most of the time. However, if more state and observation sequences are generated, using the same HMM, then the likelihood for these observation sequences is higher with the estimates obtained by the LSE algorithm than with the ones from the B-W algorithm the majority of the time; furthermore, the likelihood with the



(a) a view from a-axis

(b) a view from b-axis

Figure 9. The distribution of likelihood $P\left(o^{1,20} \mid \hat{\theta}_{LS}\right)$ and a fixed point $\hat{\theta}_{LS}$ (pointed by an arrow), where x-, y-, and r-values are fixed as equal to those of the fixed point.

estimates obtained by the B-W algorithm are extremely close to zero also the majority of the time. Some of the simulation results are described below.

An original two-state hidden Markov model has been randomly chosen. Let this parameter set be θ. Then a set of transition and observation sequences, $s_1^{1,n}$ and $o^{1,n}$, is generated using θ, where $n = 20$ and 40. The corresponding LSE, $\hat{\theta}_{LS}$ is found, and Baum-Welch estimates with randomly chosen 30 initial estimates $\hat{\theta}_i^{(1)}$, $i = 1, 2, \ldots, 30$, are found. Among those 30 Baum-Welch estimates, the one with the largest basin, $\hat{\theta}_{BW}$ is chosen. Furthermore, a particle filter estimation $\hat{\theta}_{PF}$ is made using 10,000 particles.

Then, using the same parameter set θ, twenty more sets of transition and observation sequences of length n are generated, and the corresponding likelihood $P(o_j^{1,n} \mid \hat{\theta})$, $j = 1, 2, \ldots, 20$, are found, where $\hat{\theta} = \{\hat{\theta}_{LS}, \hat{\theta}_{PF}, \hat{\theta}_{BW}\}$ are the estimates obtained from the originally generated observation sequence.

Figures 10 and 11 are one example of the results. Three likelihood values on the

leftmost vertical line is the likelihood found for the original observation sequence $o^{1,n}$. As you can see, in general, except for the likelihood with the original observation sequence, the LSE estimate often has higher likelihood compared to the B-W estimate, while the likelihood for the Baum-Welch estimates often stays almost zero. Meanwhile, the likelihood given the particle filter estimate stays almost the same for any sequences when the sequence length n is 20. When $n = 40$, the particle filter estimate happens to be very close to the LSE estimate; so, the likelihood values given these two estimates are almost identical.

The behavior of these three estimates varies depending on the true parameter values of HMM; however, this example is one of the typical cases.

As for the example shown in Figures 10 and 11, the true parameter set used is $\theta = (a, b, x, y, r) = (0.22, 0.14, 0.39, 0.37, 0.90)$. In the experiment of length-twenty sequences, the original state sequence is

$$s^{1,20} = (0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0)$$

and the original observation sequence used to have the estimates is

$$o^{1,20} = (1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1);$$

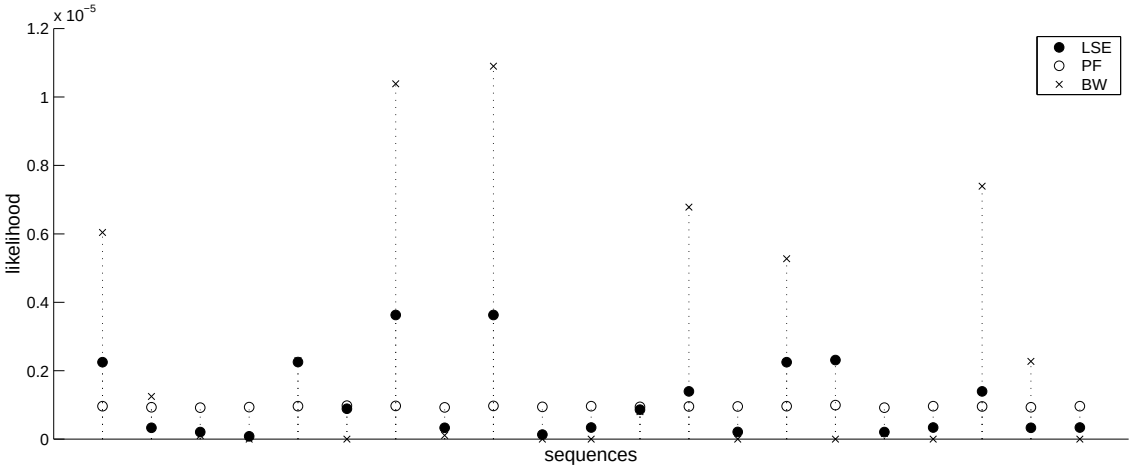
the LSE estimate $\hat{\theta}_{LS}$, the particle filter estimate $\hat{\theta}_{PF}$, and the Baum-Welch estimate $\hat{\theta}_{BW}$ are

$$\hat{\theta}_{LS} \approx (0.643, 0.323, 0.367, 0.589, 0.511),$$

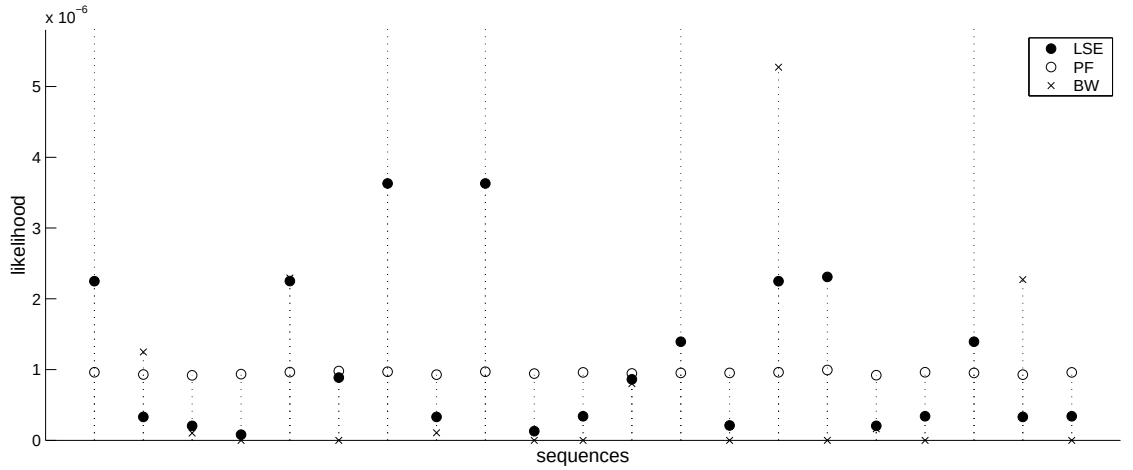
$$\hat{\theta}_{PF} \approx (0.642, 0.318, 0.479, 0.467, 0.500), \quad \text{and}$$

$$\hat{\theta}_{BW} \approx (0.497, 0, 0.549, 1, 0);$$

and the likelihood values with the original observation sequence are $P(o^{1,20} \mid \hat{\theta}_{LS}) \approx 2.249^{-6}$, $P(o^{1,20} \mid \hat{\theta}_{PF}) \approx 0.962^{-6}$, and $P(o^{1,20} \mid \hat{\theta}_{BW}) \approx 6.039^{-6}$. The average likelihood over the

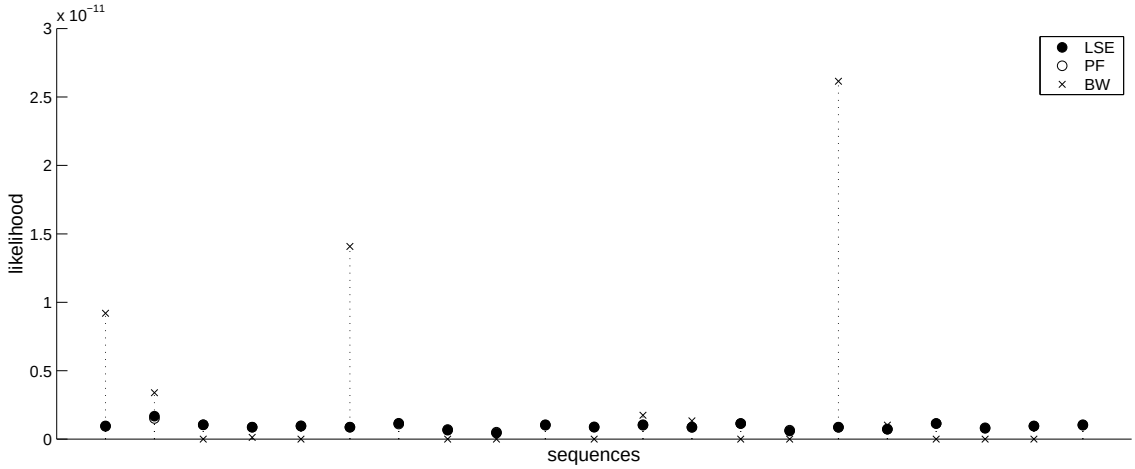


(a) Entire view

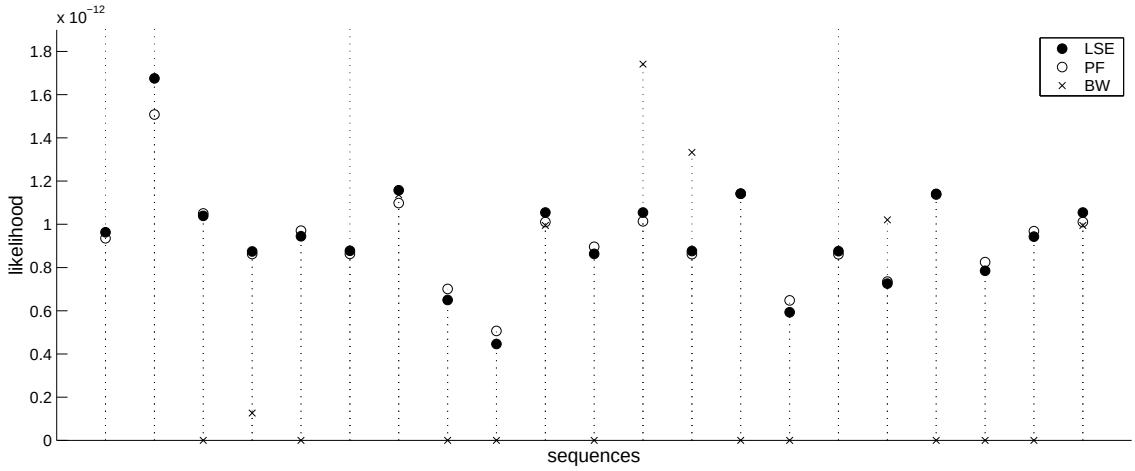


(b) Zoomed-in view

Figure 10. Likelihood change with state space size 2 (sequence length = 20)



(a) Entire view



(b) Zoomed-in view

Figure 11. Likelihood change with state space size 2 (sequence length = 40)

twenty additional observation sequences are approximately 1.073^{-6}, 0.952^{-6}, and 2.386^{-6} for $\hat{\theta}_{LS}$, $\hat{\theta}_{PF}$, and $\hat{\theta}_{BW}$, respectively.

In the experiment implemented with the sequence length fixed as 40, the original state sequence generated (from the same θ above) is

$$s^{1,40} = (0,1,0,1,0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,0,$$
$$1,0,1,0,0,1,0,1,0,1,0,0,1,1,0,1,0,1,1,0)$$

and the original observation sequence is

$$o^{1,40} = (1,0,1,0,1,1,0,1,1,0,0,1,1,0,0,1,0,0,1,1,$$
$$0,1,1,0,1,0,1,0,0,1,1,0,0,1,0,0,1,1,1,0).$$

With the same $\theta = (a, b, x, y, r) = (0.22, 0.14, 0.39, 0.37, 0.90)$, we got the LSE estimate $\hat{\theta}_{LS}$, the particle filter estimate $\hat{\theta}_{PF}$, and the Baum-Welch estimate $\hat{\theta}_{BW}$ are

$$\hat{\theta}_{LS} \approx (0.615, 0.289, 0.486, 0.541, 0.502),$$
$$\hat{\theta}_{PF} \approx (0.641, 0.324, 0.463, 0.489, 0.461), \quad \text{and}$$
$$\hat{\theta}_{BW} \approx (0.333, 381, 1, 1, 0);$$

the likelihood values obtained for the original observation sequence are $P(o^{1,40} \mid \hat{\theta}_{LS}) \approx 0.963^{-12}$, $P(o^{1,40} \mid \hat{\theta}_{PF}) \approx 0.936^{-12}$, and $P(o^{1,40} \mid \hat{\theta}_{BW}) \approx 9.195^{-12}$. The average likelihood over the twenty additional observation sequences generated using θ are approximately 0.963^{-12}, 0.935^{-12}, and 9.195^{-12} for $\hat{\theta}_{LS}$, $\hat{\theta}_{PF}$, and $\hat{\theta}_{BW}$, respectively.

Similar results can be seen with the larger state space size. Figure 12 shows a typical

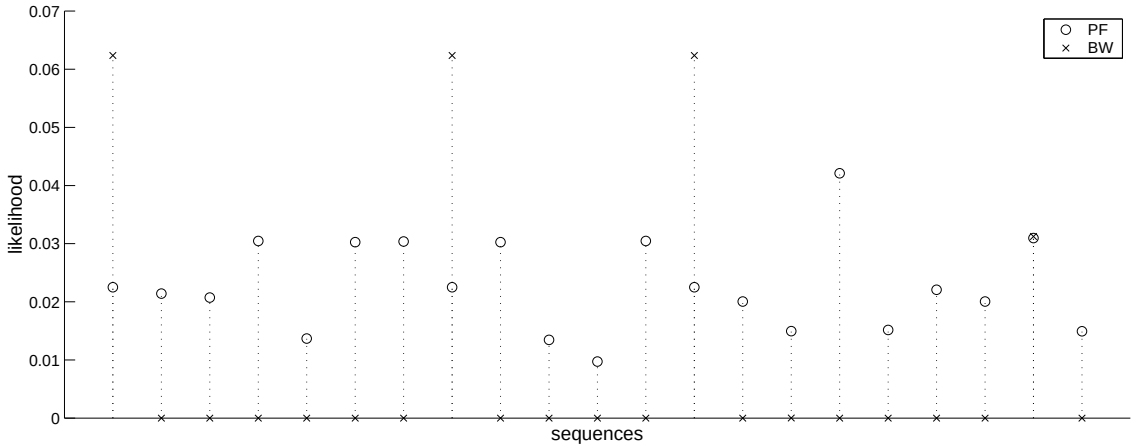


Figure 12. Likelihood change with state space size 5 (sequence length = 6)

result with $m_A = 5$. The true parameter set used is

$$A = \begin{pmatrix} 0.203 & 0.237 & 0.400 & 0.003 & 0.157 \\ 0.085 & 0.072 & 0.634 & 0.168 & 0.041 \\ 0.661 & 0.066 & 0.068 & 0.088 & 0.117 \\ 0.401 & 0.097 & 0.286 & 0.048 & 0.168 \\ 0.478 & 0.273 & 0.128 & 0.111 & 0.010 \end{pmatrix} \qquad B = \begin{pmatrix} 0.726 & 0.274 \\ 0.180 & 0.820 \\ 0.938 & 0.062 \\ 0.259 & 0.741 \\ 0.838 & 0.162 \end{pmatrix},$$

and, using this θ, a state sequence and an observation sequence are generated for estimation, which are $s^{1,6} = (1, 3, 2, 0, 2, 0)$ and $o^{1,6} = (0, 1, 0, 1, 0, 0)$. Then, a particle filter estimation and a B-W estimation are found. As for the particle filter method, 1000 particles are used. As for the B-W estimation, 30 randomly generated initial theta values are used, then the estimate with a largest basin is chosen. The estimates obtained are as shown below.

$\hat{\theta}_{PF}$:

$$\hat{A}_{PF} \approx \begin{pmatrix} 0.301 & 0.474 & 0.091 & 0.058 & 0.076 \\ 0.459 & 0.152 & 0.149 & 0.130 & 0.109 \\ 0.588 & 0.118 & 0.054 & 0.124 & 0.116 \\ 0.606 & 0.184 & 0.100 & 0.024 & 0.086 \\ 0.531 & 0.161 & 0.154 & 0.145 & 0.009 \end{pmatrix} \qquad \hat{B}_{PF} \approx \begin{pmatrix} 0.733 & 0.267 \\ 0.413 & 0.587 \\ 0.569 & 0.431 \\ 0.661 & 0.339 \\ 0.496 & 0.504 \end{pmatrix}$$

$\hat{\theta}_{BW}$:

$$\hat{A}_{BW} \approx \begin{pmatrix} 0.500 & 0.500 & 0.000 & 0.000 & 0.000 \\ 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \end{pmatrix} \qquad \hat{B}_{BW} \approx \begin{pmatrix} 1.000 & 0.000 \\ 0.000 & 1.000 \\ 1.000 & 0.000 \\ 1.000 & 0.000 \\ 1.000 & 0.000 \end{pmatrix}$$

Then, 20 more sets of state and observation sequences, $o^{1,6}(1), o^{1,6}(2), \ldots, o^{1,6}(20)$ are generated using the same parameter set θ, and the likelihood values for these sequences, $P\left(o^{1,6}(i) \mid \hat{\theta}_{PF}\right)$ and $P\left(o^{1,6}(i) \mid \hat{\theta}_{BW}\right)$, $i = 1, 2, \ldots, 20$, are obtained.

Though the B-W estimate maximizes the likelihood of the original observation sequence $P\left(o^{1,6} \mid \hat{\theta}\right)$, and though it reaches a very high likelihood occasionally also with the 20 additional observation sequences, more often it gives almost zero likelihood values for the additional sequences, basically because the estimates are on the border of the parameter space.

In order to see how often the likelihood is close to zero with the estimates from the B-W algorithm, 300 parameter sets θ_i, $i = 0, 1, \ldots, 300$, have been randomly generated, then for each θ_i, one set of state and observation sequences of length 20 is generated. Then, for each of these 300 observation sequences, say $o_i^{1,20}$, the LSE is found. Also, for each

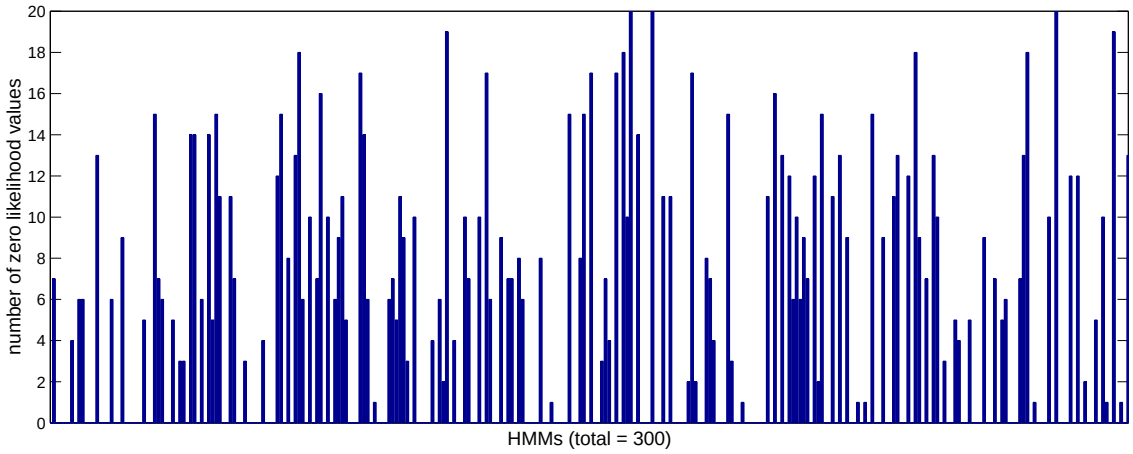


Figure 13. Frequency for B-W estimator having zero likelihood (sequence length = 20)

i, a particle filter estimate is made using 5000 particles, and one B-W estimate with the largest basin among 30 initial estimate values is chosen. Denote them as $\hat{\theta}_{LS}(i)$, $\hat{\theta}_{PF}(i)$, and $\hat{\theta}_{BW}(i)$. As expected, the likelihood given the B-W estimate, $P\left(o_i^{1,20} \mid \hat{\theta}_{BW}(i)\right)$, is greater than the likelihood with the LSE, $P\left(o_i^{1,20} \mid \hat{\theta}_{LS}(i)\right)$, for all $i = 0, 1, \ldots, 300$.

Then, 20 more sets of state and observation sequences are generated, using each θ_i. Denote these observation sequences as $o_i^{1,20}(j)$, $j = 1, 2, \ldots, 20$; which indicates θ_i is used to generate $o_i^{1,20}(1)$, $o_i^{1,20}(2)$, $\ldots$, $o_i^{1,20}(20)$.

First, the average likelihood value among $300 \cdot 20 = 6000$ of $P\left(o_i^{1,20}(j) \mid \hat{\theta}_{LS}(i)\right)$, $P\left(o_i^{1,20}(j) \mid \hat{\theta}_{PF}(i)\right)$, and $P\left(o_i^{1,20}(j) \mid \hat{\theta}_{BW}(i)\right)$ are approximately 0.001427, 0.000001, and 0.012064, respectively. So, the B-W estimates are about ten-times better than LSE estimates, and 10000-times better than particle filter estimates in this aspect. However, if we consider any likelihood less than 10^{-50} as zero, while LSE nor particle filter method never gives zero likelihood values for any of the additional observation sequences, approximately 21 percent of the time the B-W estimates give the likelihood value zero. Figure 13 shows

number of times the likelihood of an observation sequence goes to zero given the B-W estimate for each of 300 HMM with θ_i. (For example, if you see the vertical bar reaching the top of the window, the likelihood of the observation sequence given the B-W estimate is zero for all the additional 20 observation sequences for that particular HMM.) As you can see from this figure, although the B-W estimate gives a very high likelihood for some of the sequences and so the average likelihood is still higher than that of the LSE, more often, it gives very small likelihood values.

As a result, if we count the number of times one estimate gives the highest likelihood value, about 47 percent of the time the likelihood of the observation sequence given the LSE, $P\left(o_i^{1,20}(j) \mid \hat{\theta}_{LS}(i)\right)$, $i = 1, 2, \ldots, 300$ and $j = 1, 2, \ldots, 20$, is the highest among the three, followed by 29 percent by B-W estimates. Meanwhile, about 63 percent of the time, the likelihood given the corresponding B-W estimates is the lowest among the three.

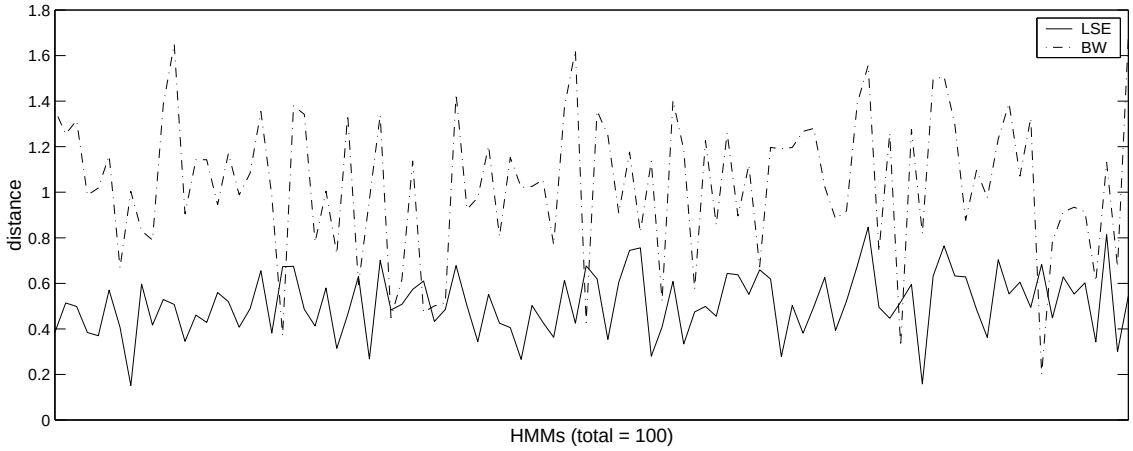


Figure 14. The Euclidean distance between the true parameters and their estimates (sequence length = 20)

A result from another kind of experiment is shown in Figure 14 agrees what we have so far. One hundred θ_i, $i = 1, 2, \ldots, 100$, are randomly chosen, then for each θ_i, a set of

state and observation sequences are generated. Using 30 randomly chosen initial estimates, B-W estimates are obtained and also LSE estimates $\hat{\theta}_{LS}(i)$ are found for each i. Then, as before, the estimates with the largest basin, $\hat{\theta}_{BW}(i)$, is chosen to be used. The Euclidean distance between the true parameter and these parameter estimates are plotted. The distance between B-W estimates and the true parameters (plotted with a dashed line) is consistently larger than that between the LSE estimates and the true parameters. We can see that maximizing the likelihood for a particular observation sequence would not necessarily result in finding a parameter estimate that is the closest among estimates obtained by other methods.

2. Jacobian of Baum-Welch Computation

Using the iterative formula shown in Section 2, some experimental results for HMM of state space size two have been obtained.

Three thousand observation sequences are randomly generated, and for each of those sequences one initial value of the estimate is randomly chosen, which is then used to find the corresponding B-W estimate. With the sequence length 20, the mean of the determinant of the Jacobian matrices was approximately $-1 \cdot 10^{-6}$ with variance approximately $1 \cdot 10^{-7}$, and the highest entry value of the Jacobian was approximately 70.

As for the eigenvalues, out of 3000 Jacobian matrices, 2491 have at least one eigenvalue that is less than 0.0001 in magnitude, among which 188 have at least one zero eigenvalue, 196 have at least one eigenvalue that is greater than one, and 471 have at least one pair of imaginary eigenvalue.

Per matrix (among five eigenvalues), the mean of the number of eigenvalues that is less than 0.0001 in magnitude is 1.857, of the number of zero eigenvalues is 0.068, of the

number of eigenvalues that is greater than one is approximately 0.074, and of the number of pairs of imaginary eigenvalues is 0.157; while the overall mean of the real-value eigenvalues is approximately 0.286.

The meaning of these numbers is as follows:

- The fact that most of the time most eigenvalues are greater than zero but less than one implies that the convergence of the algorithm is "linear"; i.e., the distance to the fixed point decreases exponentially.

- The fact that some percentage of time there is an eigenvalue larger than one suggests strongly that there are numerical "fixed points" of the B-W algorithm which represent close approaches to a saddle point along a stable manifold. As the B-W algorithm is implemented by iteration of a continuous mapping, it is evident that such saddle points exist; e.g., on the boundary of basins of attraction.

- The fact that imaginary eigenvalues exist suggest that in some numerical "fixed points" are either very slowly converging to local maxima of the likelihood or else may not be local maxima at all.

CHAPTER 5

Conclusions

The Baum-Welch algorithm is widely used for parameter estimation of hidden Markov models. Although its advantages are clear, its shortcomings, such as the "over-fitting" and the difficulty in making the algorithm online, are also discussed and some modified versions are proposed in various papers. On the other hand, the LSE estimation has largely been disregarded in the literature; we suppose mainly because of the cost of the implementation. However, experimental results presented here show that the quality of parameter estimates obtained by LSE estimation is too high for the method to be ignored entirely. The LSE estimates are simply much closer to the true parameter values than the B-W estimates most of the time.

As for the computational complexity of the LSE estimation, the algorithm introduced in this paper provides significant complexity reduction with respect to the sequence length. However, it is still exponential in the dimension of the state space. Hence, for the state space size up to four or five, we consider the LSE estimation to be feasible, while for the larger state space size, we recommend an approximation for the LSE estimation; for example, a particle filter estimation. The basic particle filter method applied here provides approximations to the LSE estimates. An improvement in both the exact LSE estimation and its approximation should be studied further.

REFERENCES

[1] D. M. Arnold, *Computing Information Rates of Finite-State Models with Application to Magnetic Recording*, Hartung Gorre, Wolfgang, 2003.

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[3] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972

[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains," *Annuals of Mathematical Statistics*, vol. 41, pp. 164–171, Feb. 1970.

[5] J. A. Bilms, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," International Computer Science Institute, Tech. Rep. ICSI-TR-97-021, April 1998.

[6] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Comput. Vision Image Understand.*, vol. 91, no. 1–2, pp. 160–187, Jul. 2003.

[7] R. I. A. Davis, B. C. Lovell, and T. Caelli, "Improved estimation of hidden Markov model parameters from multiple observation sequences," ed. R. Kasturi, D. Laurendeau, and C. Suen, in *Proc. 16th Int. Conf. Pattern Recognition*, vol. 2, Quebec, Canada, Aug. 2002, pp. 168–171.

[8] A. D. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B.*, vol. 39, no. 1, pp. 1–38, 1977.

[9] V. Digalakis, S. Tsakalidis, C. Harizakis, and L. Neumeyer, "Efficient speech recognition using subvector quantization and discrete-mixture HMMs," *Computer Speech Language*, vol. 14, no. 1, pp. 33–46, Jan. 2000.

[10] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Stat. Comput.*, vol. 10, no. 3, pp. 197–208, Jul. 2000.

[11] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[12] R. Fermandez and R. W. Picard, "Signal Processing for Recognition of Human Frustration," in *Proc. IEEE ICASSP 98*, Seattle, W.A. 1997.

[13] S. Günter and H. Bunke, "Optimizing the number of states, training iterations and Gaussians in an HMM-based handwritten word recognizer," in *Proc. 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003)*, 2003, pp. 472–476.

[14] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, Sep. 2000.

[15] F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, MA, 1998.

[16] J. Kazama, Y. Miyao, and J. Tsujii, "A maximum entropy tagger with unsupervised hidden Markov models," in *Proc. 6th Natural Language Processing Pacific Rim Symposium (NLPRS2001)*, 2001, pp. 333–340.

[17] J. Kim, K. V. Palem, and W.-F. Wong, "A framework for data prefetching using off-line training of Markovian predictors," *ICCD 2002*, pp. 340–347, 2002.

[18] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell System Tec. J.*, vol. 62, no. 4, pp.1035–1074, Apr. 1983.

[19] N. Liu and B. C. Lovell, "Gesture classification using hidden Markov models and Viterbi path counting," in *Proc. 7th Digital Image Computing: Techniques and Applications*, ed. C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, Dec. 2003, pp. 273–282.

[20] J. Ma, L. Xu, and M. I. Jordan, "Asymptotic convergence rate of the EM algorithm for Gaussian mixtures," *Neural Comput.*, vol. 12, no. 12, pp. 2881–2907, Dec. 2000.

[21] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[22] S. Müller, F. Wallhoff, F. Hülsken, and G. Rigoll, "Facial expression recognition using pseudo 3-D hidden Markov models," in *Proc. 16th Int. Conf. on Pattern Recognition*, ICPR 2002, Aug. 2002, pp. 11–15.

[23] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, pp. 47–60, Nov. 1996.

[24] N. Oliver, E. Horvitz, and A. Garg, "Layered Representations for human activity recognition," in *Proc. 4th IEEE Int. Conf. Multimodal Interfaces*, Oct. 2002, pp. 3–8.

[26] J. S. Pedersen and J. Hein, "Gene finding with a hidden Markov model of genome structure and evolution," *Bioinformatics*, vol. 19, no. 2, pp. 219-227, 2003.

[27] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[28] R. Redner and H. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

[29] P. A. Schrodt, "Pattern recognition of international crises using hidden Markov models," *Political Complexity: Nonlinear Models of Politics*, ed. Diana Richards, pp. 296–328, Ann Arbor: University of Michigan Press.

[30] H. Shu, I. L. Hetherington, and J. Glass, "Baum-Welch training for segment-based speech recognition," in *Proc. 2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, 2003, pp. 43–48.

[31] S. Tao and R. Guérin, "On-line estimation of internet path performance: An application perspective," in *Proc. IEEE INFORCOM 2004*, Mar. 2004.

[32] G. E. Tusnády and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, no. 9, pp. 849-850, 2001.

[33] W. Zucchini and P. Guttorp, "A hidden Markov model for space-time precipitation," *Water Resources Research*, vol. 27, no. 8, pp. 1917-1923, 1991.