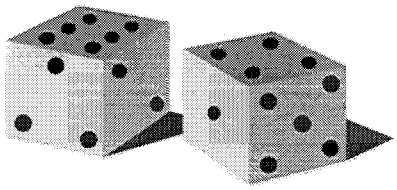
Fitting hidden semi-Markov models

John Sansom
P. J. Thomson



NIWA Technical Report 77
ISSN 1174-2631
2000

Fitting hidden semi-Markov models

John Sansom
P. J. Thomson*

***Statistics Research Associates Ltd.**
PO Box 12 649
Wellington

NIWA Technical Report 77
2000

Published by NIWA
Wellington
2000

Inquiries to:
Publication Services, NIWA,
PO Box 14-901, Wellington, New Zealand

ISSN 1174-2631
ISBN 0-478-23202-0

Citation: Sansom, J. & Thomson, P. J. 2000:
Fitting hidden semi-Markov models.
NIWA Technical Report 77. 38 p.

The National Institute of Water and Atmospheric Research
is New Zealand's leading provider
of atmospheric, marine,
and freshwater science

Visit NIWA's website at http://www.niwa.cri.nz

Contents

Abstract

Sansom, J. & Thomson, P.J. 2000: Fitting hidden semi-Markov models.
***NIWA Technical Report* 77. 38 p.**

The hidden semi-Markov model is described, its likelihood derived, and the Expectation Maximisation algorithm is applied to the likelihood to derive formulae for estimating the model's parameters. These fall into three groups. Firstly, those of the observations distributions which are assumed to be normal or can be closely approximated as a mixture of normals. The second group are those of the state dwell time distributions, and results for a number of different distributions are given with, in particular, a distribution comprising a mixture with disjoint ranges being offered as a flexible choice. Finally, the transition matrix controlling the changing of states is treated as non-parameterised with all its elements independent of each other apart from the constraint that the matrix's rows must sum to unity.

The model is extended to deal with a censored dataset in which for some observations the value is ignored and only the fact that it occurred is retained. The form of censoring dealt with is where observations below a threshold value for univariate observations are ignored or, for bivariate observations, those below or to the left of a line across the plane of the variates are ignored. Expressions for the mean and variance of the observations in the censored area, which are needed for fitting the model to censored data, are given for the normal case. Also it is shown that the model contains both the more common hidden Markov model and the conventional mixture model. The former requires the state dwell time distributions to be geometric and the latter, in addition to geometric dwell times, requires a special parameterisation of the transition matrix.

The parameter estimation formulae all include probabilities which, for a sufficiently large dataset, can be so small that they can no longer be represented within a computer. The application of suitable scaling is described and the estimation formulae recast in terms of the scaled probabilities. Some ideas regarding the practical computation of the scaled probabilities are also given. Finally, an implementation of the Viterbi algorithm is described to show how, after estimating the model parameters, a state can be attributed to each observation.

Introduction

The behaviour of a set of data may be understood by fitting a model to it. Models can take many forms but are restricted in this report to those suitable for data generated by a stochastic process which is a collection of random variables indexed by time and specified by an appropriate set of parameters. The estimation of the parameters for a particular model from the data constitutes the model fitting process. From the fitted model, inferences and predictions of the behaviour of the process can then be made.

The stochastic models considered in this report are those in which dependence from one observation to the next in the time ordered sequence can be dealt with by assuming a Markovian structure. There are many books which give details of the theory of Markov chains and Markov processes, for example, Cox & Miller (1965) and Karlin & Taylor (1975). For each observation in a Markov process, the state associated with the next observation depends only on the state of the current observation and is independent of all previous states. Thus the history of the process does not affect the next transition except through the current state. For each ordered pair of states a transition probability exists and needs to be estimated as part of the fitting procedure. Together these probabilities constitute the transition probability matrix.

A consequence of adopting a simple Markov chain structure for the observations is that the lengths of sequences of self-transitions, or the dwell times in a particular state, have a geometric distribution. A generalisation away from this can be made by fitting the observed dwell times to a chosen distribution. This modification is made to the models considered in this report and gives what is termed a semi-Markov structure.

The final aspect of the models considered in this report is the qualification *hidden*. For the data to hand it is supposed that, with the sequence of observations, there is associated an underlying but unobserved semi-Markov process whose state defines the particular distribution for each observation. Then not knowing what state was current when each observation was gathered gives rise to the hidden semi-Markov model (HSMM). Essentially, the hidden states are a way of describing the serial dependence through the data, but it would often be hoped that the states could be associated with some physically significant differences that exist.

Baum & Petrie (1966) founded the theory of hidden Markov models (HMMs) and subsequently the Baum-Welch algorithm was developed. This technique is more generally recognised as the Expectation-Maximisation (EM) algorithm as described by Dempster *et al.* (1977). Further development during the 1980s of HMMs was motivated by efforts to automate speech recognition (Levinson *et al.* 1983, Juang 1984) culminating in the review paper by Rabiner (1989). During the 1990s the books by Elliot *et al.* (1995) and MacDonald & Zucchini (1997) became available and applications in meteorology were made by Zucchini & Guttorp (1991) and Sansom (1998, 1999).

The extension to the HSMM was made by Ferguson (1980) and briefly reviewed by Rabiner (1989). However, Ferguson's important and seminal work is not easy to obtain — it is in a relatively obscure conference proceedings and is somewhat succinct — and Rabiner's review is a little too brief to cover all the issues. Thus part of the motivation for this report was to place the theory of HSMMs in a more accessible and explicit form, particularly for meteorological and environmental scientists. The remaining motivation was to extend what had been done into a more general framework which includes methods for dealing with censored datasets, HMMs, and conventional mixture models.

A formal outline of the HSMM is given and its likelihood calculated. Since the state information is missing the likelihood calculation involves the careful, and sometimes subtle, use of the EM algorithm. Subsequent sections show how the likelihood can be maximised and re-estimation formulae for the model parameters derived. Some particular cases for the dwell time distribution are considered. Various useful probabilities and the formulae for calculating them are defined (derivations are given in Appendix 1). The theory is extended to cover the application to censored datasets and it is shown that the HMM and conventional mixture models are special cases of the HSMM. It is shown why scaling is necessary in the recursive procedures used and how it can be achieved. Formulae suitable for direct implementation in a computer program are given along with the formulae for re-estimation of the observation distribution parameters. Finally the Viterbi algorithm is described.

Outline of the hidden semi-Markov model

The real-valued, continuous vector observations $\mathbf{O}_t$ $(t = 1, \ldots, T)$ are available and, without loss of generality, are considered to be normally distributed. Their time order is known and it is supposed that with each observation in the sequence a label could be associated to denote in which state the system was when the observation was taken. However, these state labels (i.e., Q_t $t = 1, \ldots, T$) are not part of the observation: they are unknown and so the states are *hidden.* Thus the system, which can be in a number of states $S_1, \ldots, S_N$ say, dwells in a state (i.e., $Q_{t+1} = Q_t$) or makes a transition between states (i.e., $Q_{t+1} \neq Q_t$) with each observation but which state (or states) is (or are) concerned is not explicitly known. Also it is assumed that the number of states is known or can be estimated by fitting the model for various N and choosing one by considering the likelihoods of the different models.

Given the state information, the observations are assumed independent with distributions

$$P(\mathbf{x} \leq \mathbf{O}_t < \mathbf{x} + d\mathbf{x} | \mathbf{Q} = \mathbf{q}) = b_{q_t}(\mathbf{x}|\Theta_{q_t})d\mathbf{x}$$

or, as only knowing Q_t is relevant to $\mathbf{O}_t$,

$$P(\mathbf{x} \leq \mathbf{O}_t < \mathbf{x} + d\mathbf{x} | Q_t = q_t) = b_{q_t}(\mathbf{x}|\Theta_{q_t})d\mathbf{x}$$

where Θ_{q_t} represents the parameters of the distribution for the observations pertaining to the state indicated by q_t. The stochastic process generating the Q_t is given by a finite Markov chain for the transitions between different states and by state dependent duration distributions for the dwell times within states. Associated with the Markov chain is a transition probability matrix $\mathbf{A} = \{a_{ij}\}$ with

$$\sum_{j=1}^{N} a_{ij} = 1 \ (i = 1, \ldots, N)$$

and since the dwell times, d, have their own distributions

$$a_{ij} = \begin{cases} P(Q_{t+1} = i | Q_t = j) & i \neq j \\ 0 & i = j \end{cases}$$

self-transitions are forbidden. If they were allowed, the model would be an HMM rather than an HSMM implying the dwell time in a particular state is geometrically distributed. However, a distribution for the dwells can be defined from the set of probabilities

$$p_d(i) = P(q_\tau = i) \ (i = 1, \ldots, N; d = 1, 2, \ldots, T; t \leq \tau \leq t + d)$$

Given that the Markov chain is irreducible — every state is accessible from every other — is positive recurrent — every state will be re-visited within finite time — and is aperiodic — no regular cycling occurs between states — then the chain has a steady state distribution π where the row vector $\pi = (\pi_1, \ldots, \pi_N)$ satisfies

$$\pi \mathbf{A} = \pi$$

and, for example, $\pi_1 T$ for large enough T is the expected number of visits to S_1. Also, it can be reasonably assumed that for a long-running process the probabilities of the state that is initially observed are given by π.

Note that the state labels Q_t are a function of the bivariate stochastic process (I_k, D_k) $(k = 1, 2, \ldots)$ where k indexes the state visits so I_k is the state visited on the kth visit and D_k is the duration of that kth visit. The relationship between the Q_t and the (I_k, D_k) is

$$\begin{aligned} q_t &= I_1 \qquad (t = 1, \ldots, D_1) \\ q_t &= I_2 \qquad (t = D_1 + 1, \ldots, D_1 + D_2) \\ \vdots &= \vdots \\ q_t &= I_R \qquad (t = T - D_R, \ldots, T) \end{aligned}$$

where R is the number of state visits up to time T. The I_k constitute a stationary finite Markov chain with probability matrix A and the probability of an observed sequence of states is

$$P(I_1 = i_1, D_1 = d_1, \ldots, I_r = i_r, D_r = d_r) = \pi_{i_1} p_{d_1}(i_1) \prod_{k=2}^{r} a_{i_{k-1} i_k} p_{d_k}(i_k)$$

which forms part of the likelihood for the complete data i.e., $(\mathbf{O}_t, Q_t)'$ $(t = 1, \ldots, T)$. The derivation of this likelihood and its maximisation are the major problems in fitting HSMMs to a dataset. A further problem dealt with in this report is the fixing of a label to each observation, i.e., finding q_t $(t = 1, \ldots, T)$ with each q_t being one of the labels $1, \ldots, N$.

Likelihood of the model

It is assumed that the available (incomplete) data are $\mathbf{O}_1, \ldots, \mathbf{O}_T$ and are taken from a process such that both $\mathbf{O}_1$ and $\mathbf{O}_T$ are state boundaries, i.e., a state started at $t = 1$ and $Q_T \neq Q_{T+1}$. Thus in terms of the generating process

$$T = D_1 + \cdots + D_R$$

Depending on how the data collection process is viewed, either or both of T and R are random variables.

In frequentist terms, the data to hand could be conceived as one realisation of many replications where R was fixed — although typically unknown — and T was random. If T is taken to be the first state boundary after a given T_0 then both T and R are random. Alternatively, T could be fixed with R random in which case the conditional distributions of the D_k become dependent. In Bayesian terms, R could be allowed to be random, with a given prior, and then T and the data generated.

In the sequel R is assumed to be a fixed, but unknown, constant. Indeed, it is R rather than T that indexes the information collection process and provides a measure of the accuracies of estimations. Thus the likelihood of the complete data is now

$$\begin{aligned} & P(I_1 = i_1, D_1 = d_1, \ldots, I_r = i_r, D_r = d_r, \mathbf{O}_1 = \mathbf{o}_1, \ldots, \mathbf{O}_T = \mathbf{o}_T | R = r) \\ = \; & P(\mathbf{O} = \mathbf{o} | \mathbf{I} = \mathbf{i}, \mathbf{D} = \mathbf{d}, R = r) P(\mathbf{I} = \mathbf{i}, \mathbf{D} = \mathbf{d} | R = r) \\ = \; & \prod_{t=1}^{T} P(\mathbf{O}_t = \mathbf{o}_t | Q_t = q_t) \pi_{i_1} p_{d_1}(i_1) \prod_{k=2}^{r} a_{i_{k-1} i_k} p_{d_k}(i_k) \end{aligned}$$

where the relationship between the Q_t and the (I_k, D_k) has been exploited. Thus, in terms of random variables the log-likelihood is

$$\log L_c = \sum_{t=1}^{T} \log b_{Q_t}(\mathbf{O}_t | \Theta_{Q_t}) + \log P(\mathbf{I}, \mathbf{D}, R)$$

where for Gaussian observations

$$\log b_{Q_t}(\mathbf{O}_t|\Theta_{Q_t}) = -\frac{1}{2}\log|\Sigma_{Q_t}| - \frac{1}{2}(\mathbf{O}_t - \mu_{Q_t})'\Sigma_{Q_t}^{-1}(\mathbf{O}_t - \mu_{Q_t})$$

and

$$\log P(\mathbf{I}, \mathbf{D}, R) = \begin{cases} \log \pi_{I_1} + \sum_{k=2}^{R} \log a_{I_{k-1}I_k} + \sum_{k=1}^{R} \log p_{D_k}(I_k) & R > 1 \\ \log \pi_{I_1} + \log p_{D_1}(I_1) & R = 1 \end{cases}$$

The first summation in the second term on the right hand side above can be rewritten as follows

$$\begin{aligned} \sum_{k=2}^{R} \log a_{I_{k-1}I_k} &= \log a_{I_1 I_2} + \log a_{I_2 I_3} + \cdots + \log a_{I_{R-1} I_R} \\ &= \log a_{12} \times (\text{No. of transitions from } S_1 \text{ to } S_2) + \cdots + \\ &\quad \log a_{N-1\,N} \times (\text{No. of transitions from } S_{N-1} \text{ to } S_N) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \log a_{ij} \times (\text{No. of transitions from } S_i \text{ to } S_j) \end{aligned}$$

Now by defining the indicator variable

$$N_t(i,j) = \begin{cases} 1 & Q_t = i, Q_{t+1} = j \\ 0 & \text{otherwise} \end{cases}$$

it can be seen that the number of transitions from S_i to S_j up to time T is

$$\sum_{t=1}^{T-1} N_t(i,j)$$

thus

$$\sum_{k=2}^{R} \log a_{I_{k-1}I_k} = \sum_{i=1}^{N}\sum_{j=1}^{N} \log a_{ij} \sum_{t=1}^{T-1} N_t(i,j)$$

The other summation can be rewritten as follows

$$\begin{aligned} \sum_{k=1}^{R} \log p_{D_k}(I_k) &= \log p_{D_1}(I_1) + \log p_{D_2}(I_2) + \cdots + \log p_{D_R}(I_R) \\ &= \log p_1(S_1) \times (\text{No. of visits to } S_1 \text{ that had duration } 1) + \cdots + \\ &\quad \log p_T(S_N) \times (\text{No. of visits to } S_N \text{ that had duration } T) \\ &= \sum_{i=1}^{N}\sum_{d=1}^{T} \log p_d(i) \times (\text{No. of visits to } S_i \text{ that had duration } d) \end{aligned}$$

Note that the summation over d was terminated at T since it is given that a state ends at T and so the maximum possible duration of a state is T. Now by defining the indicator variable

$$M_t(i,d) = \begin{cases} 1 & Q_t = i, \text{state in visit of duration } d \\ 0 & \text{otherwise} \end{cases}$$

it can be seen that the number of visits to state i that have duration d is

$$\frac{1}{d}\sum_{t=1}^{T} M_t(i,d)$$

thus

$$\sum_{k=1}^{R} \log p_{D_k}(I_k) = \sum_{i=1}^{N}\sum_{d=1}^{T} \log p_d(i) \frac{1}{d}\sum_{t=1}^{T} M_t(i,d)$$

EM algorithm

The model's log-likelihood cannot be maximised with respect to $\mathbf{A}$ etc because the data are incomplete as only the $\mathbf{O}_t$ are known and not the $(\mathbf{I}_k, \mathbf{D}_k)$. However, the expected value of the log-likelihood can be found (E-step of EM algorithm) using a "best guess" for $\mathbf{A}$ etc and then this expected value can be maximised (M-step) to form a new best guess and so on. Now, using the subscript zero to denote expectations or probabilities evaluated with respect to the true model or, in the case of the EM algorithm, an approximation to the true model then

$$\begin{aligned} E_0\{\log L_c|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\} = \\ \sum_{i=1}^{N}\sum_{t=1}^{T}\log b_i(\mathbf{O}_t|\Theta_i)P(Q_t=i|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T) \\ +E_0(\log P(\mathbf{I},\mathbf{D},R)|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T) \end{aligned}$$

where

$$\begin{aligned} &E_0\{\log P(\mathbf{I},\mathbf{D},R)|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\} \\ &\quad=\sum_{i=1}^{N}\log\pi_i P(I_1=i|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T) \\ &\quad+E_0\left\{\sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\log a_{ij}\sum_{t=1}^{T-1}N_t(i,j)|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\right\} \\ &\quad+E_0\left\{\sum_{i=1}^{N}\sum_{d=1}^{T}\log p_d(i)\frac{1}{d}\sum_{t=1}^{T}M_t(i,d)|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\right\} \end{aligned}$$

Now define

$$\begin{aligned} \gamma_t(i,j) &= E_0\{N_t(i,j)|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\} \\ &= P_0(Q_t=i,Q_{t+1}=j|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T) \end{aligned}$$

and

$$\begin{aligned} \delta_t(i,d) &= E_0\{M_t(i,d)|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\} \\ &= P_0(Q_t=i,\text{state in visit of duration } d|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T) \end{aligned}$$

and

$$\gamma_t(i)=P(Q_t=i|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T)$$

It can be noted that

$$\begin{aligned} \sum_{i=1}^{N}\gamma_t(i,j) &= \gamma_{t+1}(j) \\ \sum_{j=1}^{N}\gamma_t(i,j) &= \gamma_t(i) \\ \sum_{i=1}^{N}\gamma_t(i) &= 1 \end{aligned}$$

Thus

$$
\begin{aligned}
&E_0\{\log P(\mathbf{I},\mathbf{D},R)|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\}\\
&\quad=\sum_{i=1}^{N}\log\pi_i P(Q_1=i|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T)\\
&\qquad+\sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\log a_{ij}\sum_{t=1}^{T-1}\gamma_t(i,j)\\
&\qquad+\sum_{i=1}^{N}\sum_{d=1}^{T}\log p_d(i)\frac{1}{d}\sum_{t=1}^{T}\delta_t(i,d)\\
&\quad=\sum_{i=1}^{N}\gamma_1(i)\log\pi_i\\
&\qquad+\sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\left(\sum_{t=1}^{T-1}\gamma_t(i,j)\right)\log a_{ij}\\
&\qquad+\sum_{i=1}^{N}\sum_{d=1}^{T}\left(\frac{1}{d}\sum_{t=1}^{T}\delta_t(i,d)\right)\log p_d(i)
\end{aligned}
$$

Finally

$$
\begin{aligned}
&E_0\{\log L_c|\mathbf{O}_1,\ldots,\mathbf{O}_T,Q_{T+1}\neq Q_T\}=\\
&\qquad\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_t(i)\left(-\frac{1}{2}\log|\Sigma_i|-\frac{1}{2}(\mathbf{O}_t-\mu_i)\Sigma_i^{-1}(\mathbf{O}_t-\mu_i)\right)\\
&\qquad+\sum_{i=1}^{N}\gamma_1(i)\log\pi_i+\sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\left(\sum_{t=1}^{T-1}\gamma_t(i,j)\right)\log a_{ij}\\
&\qquad+\sum_{i=1}^{N}\sum_{d=1}^{T}\left(\frac{1}{d}\sum_{t=1}^{T}\delta_t(i,d)\right)\log p_d(i)
\end{aligned}
\tag{1}
$$

and it is this expression that must be maximised with respect to the parameters.

The EM algorithm is applied as follows.

1. Guess some initial values for the Θ_i, a_{ij} and $p_d(i)$.
2. Using the current values of a_{ij} and $p_d(i)$ calculate the $\gamma_t(i,j)$ and $\delta_t(i,j)$.
3. Find the Θ_i, a_{ij} and $p_d(i)$ that maximise (1) for the calculated $\gamma_t(i,j)$ and $\delta_t(i,j)$.
4. Using the new values of Θ_i, a_{ij} and $p_d(i)$ each time repeat steps 2 and 3 until convergence is obtained.

Maximisation formulae

Maximising (1) with respect to π_i and a_{ij} yields

$$\begin{aligned}\tilde{\pi}_i &= \gamma_1(i)/\sum_{i=1}^{N}\gamma_1(i) = \gamma_1(i) \quad (i = 1, \ldots, N) \\ \tilde{a}_{ij} &= \frac{\sum_{t=1}^{T-1}\gamma_t(i,j)}{\sum_{j=1}^{N}\sum_{t=1}^{T-1}\gamma_t(i,j)} \quad (i, j = 1, \ldots, N; i \neq j)\end{aligned} \tag{2}$$

This assumes that the Markov chain does not start from its stationary distribution but if the contrary were assumed then the optimisation is more difficult (Billingsley 1961) However, the term $\sum_{i=1}^{N}\gamma_1(i)\log\pi_i$ in this case is an asymptotically negligible component of (1) and can be ignored and the $\tilde{\pi}_i$ are then those that satisfy

$$\tilde{\pi}\tilde{\mathbf{A}} = \tilde{\pi}$$

Maximising (1) with respect to the parameters of the duration distributions requires the maximisation of

$$\sum_{d=1}^{T}\frac{1}{d}\Delta_d(i)\log p_d(i) \quad (i = 1, \ldots, N) \tag{3}$$

where

$$\Delta_d(i) = \sum_{t=1}^{T}\delta_t(i,d)$$

and so the specific form of $p_d(i)$ must be known to complete the maximisation. Some examples follow.

Non-parametric

Here the dwells have a discrete distribution over the range $1, \ldots, D$ and the probabilities $p_d(i)(d = 1, \ldots, D)$ must be estimated with the constraint that

$$\sum_{d=1}^{D}p_d(i) = 1$$

optimising yields

$$\tilde{p}_d(i) = \frac{\frac{1}{d}\Delta_d(i)}{\sum_{d=1}^{D}\frac{1}{d}\Delta_d(i)}$$

Geometric

Here

$$p_d(i) = (1 - p_i)^{d-1}p_i \quad (d = 1, 2, \ldots)$$

and (3) becomes

$$\sum_{d=1}^{T}\frac{1}{d}\Delta_d(i)\left(\log p_i + (d-1)\log(1-p_i)\right)$$

optimising yields

$$\tilde{p}_i = \frac{\sum_{d=1}^{T}\frac{1}{d}\Delta_d(i)}{\sum_{d=1}^{T}\Delta_d(i)}$$

Poisson

Here

$$p_d(i) = e^{-\lambda_i} \frac{\lambda_i^{d-1}}{(d-1)!} \quad (d = 1, 2, \ldots)$$

and (3) becomes

$$\sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) \left(-\lambda_i + (d-1) \log \lambda_i - \log(d-1)!\right)$$

optimising yields

$$\tilde{\lambda}_i = \frac{\sum_{d=1}^{T} \Delta_d(i)}{\sum_{d=1}^{T} \frac{1}{d} \Delta_d(i)} - 1$$

Mixture of geometrics

Here

$$p_d(i) = \phi p_i (1-p_i)^{d-1} + (1-\phi) q_i (1-q_i)^{d-1} \quad (d = 1, 2, \ldots)$$

where $0 \leq p_i, q_i < 1$ and $0 \leq \phi \leq 1$. Now (3) becomes

$$\sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) \log \left(\phi p_i (1-p_i)^{d-1} + (1-\phi) q_i (1-q_i)^{d-1} \right)$$

optimising yields the following expressions

$$\begin{aligned} \frac{1}{p_i} \sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) &= p_i \sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) \frac{(d-1)(1-p_i)^{d-2}}{p_d(i)} \\ \frac{1}{q_i} \sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) &= q_i \sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) \frac{(d-1)(1-q_i)^{d-2}}{p_d(i)} \end{aligned}$$

$$\left(\phi \frac{1-p_i}{p_i} + (1-\phi) \frac{1-q_i}{q_i} \right) \sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) = \sum_{d=1}^{T} \frac{1}{d} \Delta_d(i)(d-1)$$

The last expression can be re-arranged to give an expression for ϕ which can then be used in the first two expressions to eliminate ϕ from $p_d(i)$. These two equations can only be solved numerically for p_i and q_i but for the EM algorithm it is preferable if simple analytic solutions exist since a solution is required for every iteration of the algorithm. Thus if a mixture of geometrics was chosen simply to fit a distribution, which could not be fitted by any simple parametric form, rather than for a particular physical reason, then, another more flexible distribution which does admit analytic solutions would be more appropriate. The mixed-range distribution below is suggested as a means of providing such a distribution.

Mixed range distributions

Consider

$$p_d(i) = \phi_k p_{ik}(d; \theta_k) \quad (d \in D_k; k = 1, \ldots, m) \tag{4}$$

where the D_k are disjoint sets of positive integers

$$D_k = \{d_{k-1}, d_{k-1} + 1, \ldots, d_k - 1\} \quad (k = 1, \ldots, m)$$

with $1 = d_0 < d_1 < \cdots < d_{m-1} < T$ and $d_m = \infty$. Moreover, $0 \leq \phi_k \leq 1$ with $\sum_{k=1}^{m} \phi_k = 1$ and the $p_{ik}(d;\theta_k)$ are discrete probability functions on D_k with parameters θ_k so that

$$\sum_{d \in D_k} p_{ik}(d;\theta_k) = 1$$

Thus (4) combines together discrete distributions with disjoint ranges to make one flexible probability function; it is a mixture of distributions with disjoint ranges. For example, with $m = 2$ and

$$\begin{aligned} p_{i1}(d;\theta_1) &= p_d(i) \quad (d = 1, \ldots, D-1) \\ p_{i2}(d;\theta_2) &= (1-p_i)p_i^{d-D} \quad (d = D, \ldots) \end{aligned}$$

i.e., the first $D-1$ probabilities are "non-parametric" while the tail of the distribution is geometric. Now, in the general case here (3) becomes

$$\sum_{k=1}^{m} \sum_{d \in D_k} \frac{1}{d} \Delta_d(i) (\log \phi_k + \log p_{ik}(d;\theta_k))$$

and optimising this by differentiating with respect to ϕ_k under the constraint that $\sum_{k=1}^{m} \phi_k = 1$ yields

$$\tilde{\phi}_k = \frac{\sum_{d \in D_k} \frac{1}{d} \Delta_d(i)}{\sum_{d=1}^{T} \frac{1}{d} \Delta_d(i)} \quad (k = 1, \ldots, m)$$

and with respect to θ_k

$$\sum_{d \in D_k} \frac{1}{d} \Delta_d(i) \frac{\partial}{\partial \theta_k} p_{ik}(d;\theta_k) = 0$$

and suitable $p_{ik}(d;\theta_k)$ should be selected to ensure simple analytic solutions exist for all k. Returning to the example above

$$p_d(i) = \begin{cases} \phi p_d(i) & (d = 1, \ldots, D-1) \\ (1-\phi)(1-p_i)^{d-D} p_i & (d \geq D) \end{cases}$$

where $\sum_{d=1}^{D-1} p_d(i) = 1$ and the optimisation yields

$$\begin{aligned} \tilde{p}_d(i) &= \frac{\frac{1}{d} \Delta_d(i)}{\sum_{d=1}^{D-1} \frac{1}{d} \Delta_d(i)} \\ \tilde{p}_i &= \frac{\sum_{d=D}^{T} \frac{1}{d} \Delta_d(i)}{\sum_{d=D}^{T} (d-D+1) \frac{1}{d} \Delta_d(i)} \end{aligned}$$

with

$$\tilde{\phi} = \frac{\sum_{d=1}^{D-1} \frac{1}{d} \Delta_d(i)}{\sum_{d=1}^{T} \frac{1}{d} \Delta_d(i)}$$

An alternative would be to take $D_i = \{i\}$ $(i = 1, \ldots, D-1)$ i.e., $m = D$ rather than $m = 2$ and the ϕ_i become equivalent to the $p_d(i)$.

Probability formulae

The estimation of $\tilde{\mathbf{A}}$ and the $\tilde{p}_d(i)$ is made through $\gamma_t(i,j)$ and $\Delta_d(i)$ while the estimation of Θ_i requires $\gamma_t(i)$. These probabilities were defined in the last section but before deriving expressions for them, a few other probabilities will be defined and expressions for their evaluation given from derivations given in the appendix.

Definition of $\alpha_t(i)$

This is the probability of the occurrence of the first t observations with the last one being the last of a sequence from S_i i.e., define

$$\alpha_t(i) = P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t = i, Q_{t+1} \neq i)$$

which holds for $t = 1, \ldots, T$ but can be extended as shown for $t \leq 0$ so that $\alpha_{t-d}(j)$ is defined for all d

Result 1

$$\alpha_t(i) = \begin{cases} \sum_{j=1}^{N} \sum_{d=1}^{\infty} \alpha_{t-d}(j) a_{ji} p_d(i) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s) & 0 < t \leq T \\ \pi_i & t = 0 \\ 0 & t < 0 \end{cases} \tag{5}$$

where π is the stationary distribution of the Markov chain given by $\pi \mathbf{A} = \pi$. Also it can be noted that

$$\sum_{j=1}^{N} \alpha_0(j) a_{ji} = \sum_{j=1}^{N} \pi_j a_{ji} = \pi_i$$

Definition of $\alpha_t^*(i)$

This is the probability of the occurrence of the first t observations with a change to S_i starting at $t+1$ i.e., define

$$\alpha_t^*(i) = P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t \neq i, Q_{t+1} = i)$$

which holds for $t = 1, \ldots, T-1$ but not for $t = T$ since Q_{T+1} is not available. As for $\alpha_t(i)$ its range can be extended as shown for $t \leq 0$ so that $\alpha_{t-d}^*(j)$ is defined for all d

Result 2

$$\alpha_t^*(i) = \begin{cases} \sum_{j=1}^{N} \sum_{d=1}^{\infty} \alpha_{t-d}^*(j) a_{ji} p_d(j) \prod_{s=t-d+1}^{t} b_j(\mathbf{O}_s) & 0 < t \leq T \\ \pi_i & t = 0 \\ 0 & t < 0 \end{cases} \tag{6}$$

The following relationships hold between $\alpha_t(i)$ and $\alpha_t^*(i)$

Result 3

$$\alpha_t^*(i) = \sum_{j=1}^{N} \alpha_t(j) a_{ji} \tag{7}$$

$$\alpha_t(i) = \sum_{d=1}^{\infty} \alpha_{t-d}^*(i) p_d(i) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s) \tag{8}$$

Definition of $\beta_t(i)$

This is the probability of the occurrence of the observations from $t+1$ to T given that S_i ended at t i.e., define

$$\beta_t(i) = P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | Q_t = i, Q_{t+1} \neq i)$$

which holds for $t = 1, \ldots, T-1$ but not for $t = T$ since O_{T+1} is not available but its range can be extended as shown for $t \geq T$ so that $\beta_{t+d}(j)$ is defined for all d

Result 4

$$\beta_t(i) = \begin{cases} 0 & t > T \\ 1 & t = T \\ \sum_{j=1}^{N} \sum_{d=1}^{\infty} a_{ij}\, \beta_{t+d}(j) p_d(j) \prod_{s=t+1}^{t+d} b_j(\mathbf{O}_s) & 1 \leq t < T \end{cases} \tag{9}$$

Definition of $\beta_t^*(i)$

This is the probability, given that S_i began at $t+1$, of the occurrence of the observations from $t+1$ to T i.e., define

$$\beta_t^*(i) = P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | Q_t \neq i, Q_{t+1} = i)$$

which holds for $0 \leq t < T$ and as for $\beta_t(i)$ its range can be extended as shown for $t \geq T$ so that $\beta_{t+d}^*(j)$ is defined for all d

Result 5

$$\beta_t^*(i) = \begin{cases} 0 & t > T \\ 1 & t = T \\ \sum_{j=1}^{N} \sum_{d=1}^{\infty} \beta_{t+d}^*(j) a_{ij}\, p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s) & 0 \leq t < T \end{cases} \tag{10}$$

The following relationships hold between $\beta_t(i)$ and $\beta_t^*(i)$

Result 6

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, \beta_t^*(j) \tag{11}$$

$$\beta_t^*(i) = \sum_{d=1}^{\infty} \beta_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s) \tag{12}$$

Definition of $\gamma_t(i)$

This is the probability that, given all the observations, the system was in S_i at time t i.e., define

$$\gamma_t(i) = P(Q_t = i | \mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1})$$

which is required for $1 \leq t \leq T$.

Result 7

$$\gamma_t(i) = \left(\sum_{\tau=0}^{t-1} \alpha_\tau^*(i) \beta_\tau^*(i) - \sum_{\tau=1}^{t-1} \alpha_\tau(i) \beta_\tau(i) \right) \Big/ \sum_{i=1}^{N} \alpha_T(i) \tag{13}$$

where the second term in the denominator is interpreted as zero when $t = 1$. Alternatively

Result 8

$$\gamma_t(i) = \begin{cases} \alpha_0^*(i)\beta_0^*(i) \Big/ \sum_{i=1}^N \alpha_T(i) & t = 1 \\ \gamma_{t-1}(i) + \left(\alpha_{t-1}^*(i)\beta_{t-1}^*(i) - \alpha_{t-1}(i)\beta_{t-1}(i)\right) \Big/ \sum_{i=1}^N \alpha_T(i) & 2 \leq t \leq T \end{cases} \tag{14}$$

Definition of $\gamma_t(i,j)$

This is the probability that, given all the observations, the system was in S_i until time t and and in S_j starting at time $t+1$ i.e., define

$$\gamma_t(i,j) = P(Q_t = i, Q_{t+1} = j | \mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1})$$

which holds for $t = 1, \ldots, T-1$ but not for $t = T$ since O_{T+1} is not available.

Result 9

$$\gamma_t(i,j) = \begin{cases} \alpha_t(i) a_{ij} \beta_t^*(j) \Big/ \sum_{i=1}^N \alpha_T(i) & i \neq j \\ \gamma_t(i) - \sum_{\substack{j=1 \\ j \neq i}}^N \gamma_t(i,j) & i = j \end{cases} \tag{15}$$

although a value for when $i = j$ is not generally required.

Definition of $\Delta_d(i)$

This is based on the probability that, given all the observations, the system was in S_i at time t and was in a visit of length d i.e., define

$$\delta_t(i,d) = P(Q_t = i, \text{state in visit of duration } d | \mathbf{O}_1, \ldots, \mathbf{O}_T, Q_{T+1} \neq Q_T)$$

which holds for $t = 1, \ldots, T$, but, rather than $\delta_t(i,d)$, it is its sum over t i.e., $\Delta_d(i)$ that is required.

Result 10

$$\Delta_d(i) = d \sum_{t=0}^T \beta_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s) \alpha_t^*(i) \Big/ \sum_{i=1}^N \alpha_T(i) \tag{16}$$

Censoring and truncation

In a censored dataset the values taken at an observation are ignored but not the fact that an observation took place at a certain time. In a truncated dataset not only are the values taken at an observation ignored, the occurrence of the observation is also ignored. It does not seem possible to deal with truncation within the models being considered since the times t are essential for recovering the hidden Markov structure. If the number of hidden time points between observations is not known, then little can be recovered except when successive states are independent and the time ordering of the data is immaterial. Consider the censoring case

where $\mathbf{O}_t$ is recorded only if $\mathbf{a}'\mathbf{O}_t > c$ for some known vector $\mathbf{a}$ and scalar c. Thus rather than $\mathbf{O}_1, \ldots, \mathbf{O}_T$ the observations are

$$\mathbf{O}_t^+ = \begin{cases} \mathbf{O}_t & (\mathbf{a}'\mathbf{O}_t > c) \\ \text{NA} & (\mathbf{a}'\mathbf{O}_t \leq c) \end{cases}$$

where NA denotes "not available". A replacement for (1) is required given $\mathbf{O}_1^+, \ldots, \mathbf{O}_T^+$ and the fact that a state ends at T and

$$\begin{aligned} &E_0\{\log L_c | \mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T\} = \\ &\qquad \sum_{i=1}^{N}\sum_{t=1}^{T} E_0\{\log b_i(\mathbf{O}_t|\Theta_i)|Q_t = i, \mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T\} \\ &\qquad\qquad \times P_0(Q_t = i|\mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T) \\ &\qquad + E_0\{\log P(\mathbf{I}, \mathbf{D}, R)|\mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T\} \end{aligned}$$

and from Sansom & Thomson (1998)

$$\begin{aligned} &E_0\{\log b_i(\mathbf{O}_t|\Theta_i)|Q_t = i, \mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T\} \\ &\qquad = \begin{cases} -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\text{tr}\left\{\Sigma_i^{-1}(\mathbf{O}_t - \mu_i)(\mathbf{O}_t - \mu_i)'\right\} & (\mathbf{a}'\mathbf{O}_t > c) \\ -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\text{tr}\left\{\Sigma_i^{-1}(\Sigma_{ic} + (\mu_{ic} - \mu_i)(\mu_{ic} - \mu_i)')\right\} & (\mathbf{a}'\mathbf{O}_t \leq c) \end{cases} \end{aligned}$$

where

$$\begin{aligned} \mu_{ic} &= E_0\{\mathbf{O}_t|Q_t = i, \mathbf{a}'\mathbf{O}_t \leq c\} \\ \Sigma_{ic} &= \text{Var}_0(\mathbf{O}_t|Q_t = i, \mathbf{a}'\mathbf{O}_t \leq c) \end{aligned}$$

The evaluation of these for the univariate and bivariate cases is through standard results available from Johnson & Kotz (1972), although for the bivariate case when the censoring line is neither horizontal nor vertical the data must be rotated before the standard results can be used. The standard results and their development for the rotated case are given by Results 11 and 12 in Appendix 1. The required replacement for (1) is

$$\begin{aligned} &E_0\{\log L_c|\mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T\} \\ &\quad = \sum_{i=1}^{N}\sum_{t}^{+} \gamma_t^+(i)\left(-\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\text{tr}\left\{\Sigma_i^{-1}(\mathbf{O}_t - \mu_i)(\mathbf{O}_t - \mu_i)'\right\}\right) \\ &\quad + \sum_{i=1}^{N}\sum_{t}^{-} \gamma_t^+(i)\left(-\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\text{tr}\left\{\Sigma_i^{-1}(\Sigma_{ic} + (\mu_{ic} - \mu_i)(\mu_{ic} - \mu_i)')\right\}\right) \\ &\quad + \sum_{i=1}^{N}\gamma_1^+(i)\log\pi_i + \sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{j=1}^{N}\left(\sum_{t=1}^{T-1}\gamma_t^+(i,j)\right)\log a_{ij} \\ &\quad + \sum_{i=1}^{N}\sum_{d=1}^{T}\left(\frac{1}{d}\sum_{t=1}^{T}\delta_t^+(i,d)\right)\log p_d(i) \end{aligned}$$

and it is this that needs to be maximised with respect to the parameters. In this expression $\sum_t^+$ denotes the summation over those t where $\mathbf{a}'\mathbf{O}_t > c$ and $\sum_t^-$ that over those where $\mathbf{a}'\mathbf{O}_t \leq c$. Also

$$\begin{aligned} \gamma_t^+(i) &= P_0(Q_t = i|\mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T) \\ \gamma_t^+(i,j) &= P_0(Q_t = i, Q_{t+1} = j|\mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T) \\ \delta_t^+(i,d) &= P_0(Q_t = i, \text{state in visit of duration } d|\mathbf{O}_1^+, \ldots, \mathbf{O}_T^+, Q_{T+1} \neq Q_T) \end{aligned}$$

which are evaluated in exactly the same way as $\gamma_t(i), \gamma_t(i,j)$ and $\delta_t(i,d)$ but with the $\mathbf{O}_t$ replaced by $\mathbf{O}_t^+$. Indeed, this is not limited to just the $\mathbf{O}_t^+$ and these probabilities could be evaluated for the $\mathbf{O}_t$ replaced by $\tilde{\mathbf{O}}_t = g(\mathbf{O}_t)$ where g is some arbitrary transformation. The key property is that the $\mathbf{O}_t^+$ or, equivalently, the $g(\mathbf{O}_t)$ should be conditionally independent given knowledge of the states, Q_t. Given $Q_t = i$, then $\mathbf{O}_t$ has the mixed distribution

$$\begin{aligned} P(\mathrm{x} \leq \mathbf{O}_t^+ < \mathrm{x} + d\mathrm{x} | Q_t = i) &= b_i(\mathrm{x})d\mathrm{x} \quad (\mathbf{a}'\mathrm{x} > c) \\ P(\mathbf{O}_t^+ = \mathrm{NA} | Q_t = i) &= B_i^+(c) \end{aligned}$$

where $B_i^+(\mathrm{x})$ is the cumulative distribution function of $\mathbf{a}'\mathbf{O}_t$ given $Q_t = i$ i.e., $P(\mathbf{a}'\mathbf{O}_t \leq \mathrm{x} | Q_t = i)$. Thus the $\gamma_t^+(i), \gamma_t^+(i,j)$ and $\delta_t^+(i,d)$ and the analogous forms of the $\alpha_t(i), \beta_t(i), \alpha_t^*(i)$ and $\beta_t^*(i)$ are given by Results 1–10 with the $b_i(\mathbf{O}_s)$ replaced by $B_i^+(c)$ whenever $\mathbf{a}'\mathbf{O}_s \leq c$.

Reduced models

Hidden Markov model

In an HMM the dwell times are not given an explicit distribution as they are in the *semi*-Markov case rather they are implicitly geometric with parameters equal to the elements on the diagonal of $\mathbf{A}$, the transition matrix. Now, is the fitting of an HSMM with geometric dwell times equivalent to the fitting of an HMM to the same data?

Let the $p_d(i)$ be geometric with parameter $\tilde{a}_{ii}$ $(0 < \tilde{a}_{ii} \leq 1)$ i.e.,

$$p_d(i) = \tilde{a}_{ii}^{d-1}(1 - \tilde{a}_{ii}) \quad (d = 1, 2, \ldots)$$

and define

$$\tilde{a}_{ij} = (1 - \tilde{a}_{ii})a_{ij} \quad (i \neq j, 1 \leq i, j \leq N)$$

Then the matrix $\tilde{\mathbf{A}}$ with typical element $\tilde{a}_{ij}$ $(1 \leq i, j \leq N)$ is the transition probability matrix of a finite Markov chain since

$$\tilde{a}_{ij} \geq 0, \quad \sum_{j=1}^{N} \tilde{a}_{ij} = \tilde{a}_{ii} + (1 - \tilde{a}_{ii}) \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} = 1$$

Moreover, if the stationary distribution for $\tilde{\mathbf{A}}$ is $\tilde{\pi}$ then a typical element of $\tilde{\pi}$ is

$$\tilde{\pi}_i = c\frac{\pi_i}{1 - \tilde{a}_{ii}} \quad (i = 1, \ldots, N)$$

since substitution in $\sum_{j=1}^{N} \tilde{\pi}_j \tilde{a}_{ji}$ gives

$$\begin{aligned} \sum_{j=1}^{N} \tilde{\pi}_j \tilde{a}_{ji} &= c\left(\sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_j a_{ji} + \pi_i \frac{\tilde{a}_{ii}}{1 - \tilde{a}_{ii}} \right) \\ &= c\pi_i \left(1 + \frac{\tilde{a}_{ii}}{1 - \tilde{a}_{ii}} \right) \\ &= c\frac{\pi_i}{1 - \tilde{a}_{ii}} \\ &= \tilde{\pi}_i \end{aligned}$$

which is true since $\tilde{\pi}\tilde{\mathbf{A}} = \tilde{\pi}$. Now, summing both sides of the expression for $\tilde{\pi}_i$ over i gives

$$c = \left(\sum_{i=1}^{N} \frac{\pi_i}{1 - \tilde{a}_{ii}} \right)^{-1}$$

since $\sum_{i=1}^{N} \tilde{\pi}_i = 1$. Now consider the probability of a typical path given by the event

$$\begin{aligned} C_p \;=\; & \{S_{i_1} \text{ starts at } t = 1 \text{ and ends at } t = d_1, \\ & S_{i_2} \text{ starts at } t = d_1 + 1 \text{ and ends at } t = d_1 + d_2, \\ & \cdots \\ & S_{i_r} \text{ starts at } t = d_1 + \cdots + d_{r-1} + 1 \text{ and ends at } t = d_1 + \cdots + d_r\} \end{aligned}$$

The probability of this event in terms of the HSMM with geometric dwell times is

$$\begin{aligned} P(C_p) &= P(I_1 = i_1, D_1 = d_1, I_2 = i_2, D_2 = d_2, \ldots, I_r = i_r, D_r = d_r) \\ &= \pi_{i_1} p_{d_1}(i_1) a_{i_1 i_2} p_{d_2}(i_2) \ldots a_{i_{r-1} i_r} p_{d_r}(i_r) \\ &= \pi_{i_1} \tilde{a}_{i_1 i_1}^{d_1 - 1} \tilde{a}_{i_1 i_2} \tilde{a}_{i_1 i_2}^{d_2 - 1} \ldots \tilde{a}_{i_r i_r}^{d_r - 1} (1 - \tilde{a}_{i_r i_r}) \end{aligned}$$

since $a_{i_1 i_2}(1 - \tilde{a}_{i_1 i_1}) = \tilde{a}_{i_1 i_2}$. However, set $t_k = d_1 + \cdots + d_k$ $(k = 1, \ldots, r)$ then in terms of a HMM with transition probability matrix $\tilde{\mathbf{A}}$ the probability of event C_p is

$$\begin{aligned} \tilde{P}(C_p) &= P(Q_0 \neq i_1, Q_1 = Q_2 = \cdots = Q_{t_1} = i_1, Q_{t_1+1} = \cdots = Q_{t_2} = i_2 \\ &\qquad \ldots, Q_{t_{r-1}+1} = \cdots = Q_{t_r} = i_r, Q_{t_r+1} \neq i_r) \\ &= \sum_{\substack{j=1 \\ j \neq i_1}}^{N} \tilde{\pi}_j \tilde{a}_{j i_1} \tilde{a}_{i_1 i_1}^{d_1 - 1} \tilde{a}_{i_1 i_2} \tilde{a}_{i_1 i_2}^{d_2 - 1} \ldots \tilde{a}_{i_r i_r}^{d_r - 1} (1 - \tilde{a}_{i_r i_r}) \\ &= c \pi_{i_1} \tilde{a}_{i_1 i_1}^{d_1 - 1} \tilde{a}_{i_1 i_2} \tilde{a}_{i_1 i_2}^{d_2 - 1} \ldots \tilde{a}_{i_r i_r}^{d_r - 1} (1 - \tilde{a}_{i_r i_r}) \\ &= c P(C_p) \end{aligned}$$

where it has been assumed that the HMM is stationary. Exact equivalence can be acheived if the HMM is allowed to have an initial distribution π rather than $\tilde{\pi}$. In either case the two probabilities agree up to a constant of proportionality due to the end points which are asymptotically neglible. These probabilities express that part of each model's likelihood resulting from the Markovian structure, thus by setting the dwell distributions to be geometric the HSMM fitting will fit a HMM with transition probability matrix $\tilde{\mathbf{A}}$.

Note also that the conventional HMM, in which a state does not necessarily start at $t = 1$ nor end at $t = T$, can be fitted. In this case

$$\begin{aligned} \tilde{P}(C_p) &= P(Q_1 = \cdots = Q_{t_1} = i_1, Q_{t_1+1} = \cdots = Q_{t_2} = i_2 \\ &\qquad \ldots, Q_{t_{r-1}+1} = \cdots = Q_{t_r} = i_r) \\ &= \tilde{\pi}_{i_1} \tilde{a}_{i_1 i_1}^{d_1 - 1} \tilde{a}_{i_1 i_2} \tilde{a}_{i_1 i_2}^{d_2 - 1} \ldots \tilde{a}_{i_r i_r}^{d_r - 1} \\ &= \frac{c}{(1 - \tilde{a}_{i_1 i_1})(1 - \tilde{a}_{i_r i_r})} P(C_p) \end{aligned}$$

and the end dependent factor would need to be built into the various recursions.

Independent states

Independent states occur when there is no Markovian structure and the time ordering of the data is immaterial. Within the context of HSMMs, this can be achieved through further reduction from an HMM by restricting the transition matrix such that its elements are

$$\tilde{a}_{ij} = P(Q_t = i | Q_{t-1} = j) = P(Q_t = i) = \tilde{\pi}_i$$

i.e., at any time the probability that the system is in a certain state does not even depend on the state of the prior observation. In terms of Sansom & Thomson (1998) it should be noted that $\tilde{\pi}_i$ is equivalent to their α_i i.e., the proportional representation of state i.

In the section on maximisation formulae (*see* p. 12) parameterisation of the dwell time distributions (i.e., the $p_d(i)$ of (1)) was illustrated with some example parametric forms. In order to fit a model with independent states it is necessary to also allow for parameterisation of the transition matrix (i.e., the a_{ij} of (1)). Thus, (1) would need to be optimised with the a_{ij} as the $\tilde{a}_{ij}$ defined above and with geometric $p_d(i)$s which have $\tilde{\pi}_i$ as their parameter.

Scaling

The probabilities $\alpha_t(i), \alpha_t^*(i), \beta_t(i)$, and $\beta_t^*(i)$ all become small as t increases, to the extent that they become too small to be represented within a computer. However, the problem can be circumvented by scaling them all, together with the probability of the observation (i.e., $b_i(\mathbf{O}_s)$), using one set of scaling parameters — k_t. Scaling is required only for $1 \leq t \leq T$, thus for other values of t k_t can be taken to be unity i.e., $k_t = 1,\ t \leq 0, t > T$. Let $\hat{\alpha}_t(i)$, the scaled version of $\alpha_t(i)$, be such that $\sum_{i=1}^{N} \hat{\alpha}_t(i) = 1$ and suppose

$$\hat{\alpha}_t(i) = \prod_{\tau=-\infty}^{t} k_\tau \alpha_t(i)$$

then

$$\prod_{\tau=-\infty}^{t} k_\tau = 1 \Big/ \sum_{i=1}^{N} \alpha_t(i)$$

which, in particular, gives

$$k_0 = 1 \Big/ \sum_{i=1}^{N} \alpha_0(i) = 1 \Big/ \sum_{i=1}^{N} \pi_i = 1$$

which is consistent with both the definition of $\alpha_0(i)$ and k_0. Also let

$$\hat{b}_i(\mathbf{O}_s) = k_s b_i(\mathbf{O}_s)$$

Multiply (5) through by $\prod_{\tau=-\infty}^{t} k_\tau$ to obtain

$$\prod_{\tau=-\infty}^{t} k_\tau \alpha_t(i) = \sum_{j=1}^{N} \sum_{d=1}^{\infty} \prod_{\tau=-\infty}^{t-d} k_\tau \alpha_{t-d}(j) a_{ji} p_d(i) \prod_{s=t-d+1}^{t} k_s b_i(\mathbf{O}_s)$$

or

$$\hat{\alpha}_t(i) = \sum_{j=1}^{N} \sum_{d=1}^{\infty} \hat{\alpha}_{t-d}(j) a_{ji} p_d(i) \prod_{s=t-d+1}^{t} \hat{b}_i(\mathbf{O}_s)$$

Now let

$$\hat{\beta}_t^*(i) = \prod_{\tau=t+1}^{\infty} k_\tau \beta_t^*(i)$$

and multiply (10) through by $\prod_{\tau=t+1}^{\infty} k_\tau$ to obtain

$$\prod_{\tau=t+1}^{\infty} k_\tau \beta_t^*(i) = \sum_{j=1}^{N} \sum_{d=1}^{\infty} \prod_{\tau=t+d+1}^{\infty} k_\tau \beta_{t+d}^*(j) a_{ij}\, p_d(i) \prod_{s=t+1}^{t+d} k_s b_i(\mathbf{O}_s)$$

or

$$\hat{\beta}_t^*(i) = \sum_{j=1}^{N} \sum_{d=1}^{\infty} \hat{\beta}_{t+d}^*(j) a_{ij}\, p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s)$$

Using similar definitions for $\hat{\alpha}_t^*(i)$ and $\hat{\beta}_t^*(i)$ as for $\hat{\alpha}_t(i)$ and $\hat{\beta}_t(i)$ the following expressions follow

$$\begin{aligned} \hat{\alpha}_t^*(i) &= \sum_{j=1}^{N} \hat{\alpha}_t(j) a_{ji} \\ \hat{\beta}_t(i) &= \sum_{j=1}^{N} a_{ij}\, \hat{\beta}_t^*(j) \end{aligned}$$

Formulae using scaled probabilities

Computational formula for $\hat{\alpha}_t(i)$

From (8)

$$\hat{\alpha}_t(i) = \sum_{d=1}^{t} \hat{\alpha}_{t-d}^*(i) p_d(i) \prod_{s=t-d+1}^{t} \hat{b}_i(\mathbf{O}_s)$$

since $\hat{\alpha}_{t-d}^*(i) = 0$ for $d > t$. Each term of the summation includes a further term in the product over s, thus in a situation where $\hat{b}_i(\mathbf{O}_s)$ can be zero (i.e., the observation is impossible for the state concerned) the term that first includes a $\hat{b}_i(\mathbf{O}_s) = 0$, and all subsequent terms, will be zero. Thus

$$\hat{\alpha}_t(i) = \sum_{d=1}^{d'} \hat{\alpha}_{t-d}^*(i) p_d(i) \prod_{s=t-d+1}^{t} \hat{b}_i(\mathbf{O}_s)$$

where

$$\hat{b}_i(\mathbf{O}_{t-d+1}) = \begin{cases} 0 & d = d' \\ > 0 & 1 \le d < d' \end{cases}$$

However, if $\hat{b}_i(\mathbf{O}_s)$ can never be zero then let

$$\hat{\alpha}_t(i) = \sum_{d=1}^{t} \xi_t(i,d)$$

where

$$\xi_t(i,d) = \hat{\alpha}_{t-d}^*(i) p_d(i) \prod_{s=t-d+1}^{t} \hat{b}_i(\mathbf{O}_s)$$

or

$$\xi_{t-1}(i,d-1) = \hat{\alpha}_{t-d}^*(i) p_{d-1}(i) \prod_{s=t-d+1}^{t-1} \hat{b}_i(\mathbf{O}_s)$$

thus

$$\xi_t(i,d) = \xi_{t-1}(i,d-1)\frac{p_d(i)}{p_{d-1}(i)}\hat{b}_i(\mathbf{O}_t) \tag{17}$$

i.e., a simple recursive formula. But this is valid only for $d \geq 2$ and $t \geq 1$ since $p_0(i)$ and $\xi_0(i,d)$ are undefined, thus $\xi_t(i,1)$ is required for all t. From the definition

$$\xi_t(i,1) = \hat{\alpha}^*_{t-1}(i)p_1(i)\hat{b}_i(\mathbf{O}_t)$$

Computational formula for $\hat{\beta}_t(i)$

From (12)

$$\hat{\beta}^*_t(i) = \sum_{d=1}^{T-t}\hat{\beta}_{t+d}(i)p_d(i)\prod_{s=t+1}^{t+d}\hat{b}_i(\mathbf{O}_s)$$

since $\hat{\beta}_{t+d}(i) = 0$ for $d > T-t$. As for $\hat{\alpha}_t(i)$

$$\hat{\beta}^*_t(i) = \sum_{d=1}^{d'}\hat{\beta}_{t+d}(i)p_d(i)\prod_{s=t+1}^{t+d}\hat{b}_i(\mathbf{O}_s)$$

where

$$\hat{b}_i(\mathbf{O}_{t+d}) = \begin{cases} 0 & d = d' \\ > 0 & 1 \leq d < d' \end{cases}$$

However, if $\hat{b}_i(\mathbf{O}_s)$ can never be zero then let

$$\hat{\beta}^*_t(i) = \sum_{d=1}^{T-t}\xi_t(i,d)$$

where

$$\xi_t(i,d) = \hat{\beta}_{t+d}(i)p_d(i)\prod_{s=t+1}^{t+d}\hat{b}_i(\mathbf{O}_s)$$

or

$$\xi_{t+1}(i,d-1) = \hat{\beta}_{t+d}(i)p_{d-1}(i)\prod_{s=t+1}^{t+d-1}\hat{b}_i(\mathbf{O}_s)$$

thus

$$\xi_t(i,d) = \xi_{t+1}(i,d-1)\frac{p_d(i)}{p_{d-1}(i)}\hat{b}_i(\mathbf{O}_{t+d}) \tag{18}$$

i.e., a simple recursive formula. But this is valid only for $d \geq 2$ and $t \leq T-1$ since $p_0(i)$ and $\xi_T(i,d)$ are undefined, thus $\xi_t(i,1)$ is required for all t. From the definition

$$\xi_t(i,1) = \hat{\beta}_{t+1}(i)p_1(i)\hat{b}_i(\mathbf{O}_{t+1})$$

Formula for k_t

Suppose k_s known for $s < t$ and let

$$\check{\alpha}_t(i) = \sum_{d=1}^{t}\hat{\alpha}^*_{t-d}(i)p_d(i)b_i(\mathbf{O}_t)\prod_{s=t-d+1}^{t-1}\hat{b}_i(\mathbf{O}_s)$$

i.e.,

$$\hat{\alpha}_t(i) = k_t \breve{\alpha}_t(i)$$

and since $\sum_{i=1}^{N} \hat{\alpha}_t(i) = 1$ thus

$$k_t = 1 \Big/ \sum_{i=1}^{N} \breve{\alpha}_t(i)$$

Formula for $\gamma_t(i)$

From (13)

$$\begin{aligned}
\gamma_t(i) &= \left(\sum_{\tau=0}^{t-1} \prod_{\theta=-\infty}^{\infty} k_\theta \alpha_\tau^*(i) \beta_\tau^*(i) - \sum_{\tau=1}^{t-1} \prod_{\theta=-\infty}^{\infty} k_\theta \alpha_\tau(i) \beta_\tau(i) \right) \Big/ \sum_{i=1}^{N} \prod_{\theta=-\infty}^{\infty} k_\theta \alpha_T(i) \\
&= \left(\sum_{\tau=0}^{t-1} \prod_{\theta=-\infty}^{\tau} k_\theta \alpha_\tau^*(i) \prod_{\theta=\tau+1}^{\infty} k_\theta \beta_\tau^*(i) - \sum_{\tau=1}^{t-1} \prod_{\theta=-\infty}^{\tau} k_\theta \alpha_\tau(i) \prod_{\theta=\tau+1}^{\infty} k_\theta \beta_\tau(i) \right) \Big/ \sum_{i=1}^{N} \prod_{\theta=-\infty}^{T} k_\theta \alpha_T(i) \\
&= \left(\sum_{\tau=0}^{t-1} \hat{\alpha}_\tau^*(i) \hat{\beta}_\tau^*(i) - \sum_{\tau=1}^{t-1} \hat{\alpha}_\tau(i) \hat{\beta}_\tau(i) \right) \Big/ \sum_{i=1}^{N} \hat{\alpha}_T(i) \\
&= \sum_{\tau=0}^{t-1} \hat{\alpha}_\tau^*(i) \hat{\beta}_\tau^*(i) - \sum_{\tau=1}^{t-1} \hat{\alpha}_\tau(i) \hat{\beta}_\tau(i)
\end{aligned}$$

as $\sum_{i=1}^{N} \hat{\alpha}_T(i) = 1$. Similarly for the recursive formula for $\gamma_t(i)$ (i.e., (14))

$$\gamma_t(i) = \begin{cases} \hat{\alpha}_0^*(i) \hat{\beta}_0^*(i) & t = 1 \\ \gamma_{t-1}(i) + \hat{\alpha}_{t-1}^*(i) \hat{\beta}_{t-1}^*(i) - \hat{\alpha}_{t-1}(i) \hat{\beta}_{t-1}(i) & t = 2, \ldots, T \end{cases}$$

Formula for $\Delta_d(i)$

From (16)

$$\begin{aligned}
\Delta_d(i) &= d \sum_{t=0}^{T} \prod_{\tau=-\infty}^{\infty} k_\tau \beta_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s) \alpha_t^*(i) \Big/ \sum_{i=1}^{N} \prod_{\tau=-\infty}^{\infty} k_\tau \alpha_T(i) \\
&= d \sum_{t=0}^{T} \prod_{\tau=t+d+1}^{\infty} k_\tau \beta_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} k_s b_i(\mathbf{O}_s) \prod_{\tau=-\infty}^{t} k_\tau \alpha_t^*(i) \Big/ \sum_{i=1}^{N} \prod_{\tau=-\infty}^{T} k_\tau \alpha_T(i) \\
&= d \sum_{t=0}^{T} \hat{\beta}_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s) \hat{\alpha}_t^*(i) \Big/ \sum_{i=1}^{N} \hat{\alpha}_T(i) \\
&= d \sum_{t=0}^{T} \hat{\beta}_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s) \hat{\alpha}_t^*(i)
\end{aligned}$$

From the examples given illustrating the maximisation of (3), although $\Delta_d(i)$ is required in the non-parametric case and for $d < D - 1$ in the example given for the mixed range distribution, generally $\sum_{d=1}^{T} \Delta_d(i)$ and $\sum_{d=1}^{T} \frac{1}{d} \Delta_d(i)$ are also required. Thus

$$\sum_{d=1}^{T} \frac{1}{d} \Delta_d(i) = \sum_{d=1}^{T} \sum_{t=0}^{T} \hat{\beta}_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s) \hat{\alpha}_t^*(i)$$

$$= \sum_{t=0}^{T} \hat{\alpha}_t^*(i) \sum_{d=1}^{\infty} \hat{\beta}_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s)$$
$$= \sum_{t=0}^{T} \hat{\alpha}_t^*(i) \hat{\beta}_t^*(i)$$

Similarly

$$\sum_{d=1}^{T} \Delta_d(i) = \sum_{t=0}^{T} \sum_{d=1}^{T} d\hat{\beta}_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s) \hat{\alpha}_t^*(i)$$
$$= \sum_{t=0}^{T} \hat{\alpha}_t^*(i) \xi_t(i)$$

where

$$\xi_t(i) = \sum_{d=1}^{T-t} d\hat{\beta}_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s)$$

since $\hat{\beta}_{t+d}(i) = 0$ for $d > T - t$. Successive terms of this summation generally decrease such that it can be halted before $d = T - t$ at a point when additional terms become too small to significantly alter the overall value of the summation. However, where $\hat{b}_i(\mathbf{O}_s)$ can be zero, then as for $\hat{\alpha}_t(i)$ and $\hat{\beta}_t(i)$

$$\xi_t(i) = \sum_{d=1}^{d'} d\hat{\beta}_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} \hat{b}_i(\mathbf{O}_s)$$

where d' such that

$$\hat{b}_i(\mathbf{O}_{t+d}) = \begin{cases} 0 & d = d' \\ > 0 & 1 \leq d < d' \end{cases}$$

Formula for $\tilde{a}_{ij}$

From (2) and (15)

$$\tilde{a}_{ij} = \sum_{t=1}^{T} \alpha_t(i) a_{ij} \beta_t^*(j) \Big/ \sum_{j=1}^{N} \sum_{t=1}^{T} \alpha_t(i) a_{ij} \beta_t^*(j)$$

thus

$$\tilde{a}_{ij} = \sum_{t=1}^{T} \prod_{\tau=-\infty}^{\infty} k_\tau \alpha_t(i) a_{ij} \beta_t^*(j) \Big/ \sum_{j=1}^{N} \sum_{t=1}^{T} \prod_{\tau=-\infty}^{\infty} k_\tau \alpha_t(i) a_{ij} \beta_t^*(j)$$
$$= \sum_{t=1}^{T} \prod_{\tau=-\infty}^{t} k_\tau \alpha_t(i) a_{ij} \prod_{\tau=t+1}^{\infty} k_\tau \beta_t^*(j) \Big/ \sum_{j=1}^{N} \sum_{t=1}^{T} \prod_{\tau=-\infty}^{t} k_\tau \alpha_t(i) a_{ij} \prod_{\tau=t+1}^{\infty} k_\tau \beta_t^*(j)$$
$$= \sum_{t=1}^{T} \hat{\alpha}_t(i) a_{ij} \hat{\beta}_t^*(j) \Big/ \sum_{j=1}^{N} \sum_{t=1}^{T} \hat{\alpha}_t(i) a_{ij} \hat{\beta}_t^*(j)$$

Formula for $P(\mathbf{O}_1, \ldots, \mathbf{O}_T)$

Since

$$P(\mathbf{O}_1, \ldots, \mathbf{O}_T) = \sum_{i=1}^{N} \alpha_T(i)$$

thus

$$\begin{aligned}
\log P(\mathbf{O}_1, \ldots, \mathbf{O}_T) &= \log \sum_{i=1}^{N} \alpha_T(i) \\
&= \log \left(\sum_{i=1}^{N} \hat{\alpha}_T(i) \Big/ \prod_{t=-\infty}^{T} k_t \right) \\
&= -\log \prod_{t=-\infty}^{T} k_t \\
&= -\sum_{t=1}^{T} \log k_t
\end{aligned}$$

Formulae for re-estimation of observation distribution parameters

Formulae for re-estimation of observation distribution parameters are found through the maximisation of the first term of (1) and the re-estimation formulae were given by Sansom & Thomson (1998) with $\gamma_t(i)$ replacing their $a_i(\mathrm{x})$. This assumes that each state's observations are normally distributed but more general situations can be covered by allowing a mixture of normals for each state, i.e.,

$$b_i(\mathbf{o}_t) = \sum_{m=1}^{M} f_i(m) P_m(\mathbf{o}_t | \theta_i(m))$$

where $\mathbf{o}_t$ is the tth observation which if generated by the ith state results from a process whose output can be modelled as a mixture distribution of M components each of which has a fractional representation of $f_i(m)$ with $\sum_{m=1}^{M} f_i(m) = 1$ and a density of P_m. The parameters of these densities are represented by $\theta_i(m)$ which for present purposes have been taken to be normal and so $\theta_i(m) = \{\mu_i(m), \Sigma_i(m)\}$. Furthermore, both bivariate and univariate data may be concerned and the following notation is used:

$$\begin{aligned}
\mathbf{o}_t &= \begin{pmatrix} o_{1_t} \\ o_{2_t} \end{pmatrix} \\
\mu_i(m) &= \begin{pmatrix} \mu_{o_1 i}(m) \\ \mu_{o_2 i}(m) \end{pmatrix} \\
\Sigma_i(m) &= \begin{pmatrix} \sigma^2_{o_1 i}(m) & \sigma_{o_1 o_2 i}(m) \\ \sigma_{o_1 o_2 i}(m) & \sigma^2_{o_2 i}(m) \end{pmatrix}
\end{aligned}$$

Where

$$\gamma_{it}(m) = \gamma_t(i) \frac{f_i(m) P_m(\mathbf{o}_t | \theta_i(m))}{\sum_{m=1}^{M} f_i(m) P_m(\mathbf{o}_t | \theta_i(m))}$$

it can be shown, with the tilde indicating the new estimate, that

$$\begin{aligned}
\tilde{f}_i(m) &= \frac{\sum_{t=1}^{T} \gamma_{it}(m)}{\sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{it}(m)} \\
\tilde{\mu}_i(m) &= \frac{\sum_{t}^{+} \gamma_{it}(m) \mathbf{o}_t + C_{ic}(m) \mu_{ic}(m)}{\sum_{t=1}^{T} \gamma_{it}(m)}
\end{aligned}$$

$$\tilde{\Sigma}_i(m) = \frac{\sum_t^+ \gamma_{it}(m)(\mathbf{o}_t - \tilde{\mu}_i(m))(\mathbf{o}_t - \tilde{\mu}_i(m))' + C_{ic}(m)(\Sigma_{ic}(m) + (\mu_{ic}(m) - \tilde{\mu}_i(m))(\mu_{ic}(m) - \tilde{\mu}_i(m))')}{\sum_{t=1}^T \gamma_{it}(m)}$$

where $\sum_t^+$, $\mu_{ic}(m)$ and $\Sigma_{ic}(m)$ are as defined in the section on censoring and truncation (*see* p. 17) and

$$C_{ic}(m) = \frac{\tilde{f}_i(m)\Phi(z_m(i))}{\frac{T}{T^-}\sum_{m=1}^M \tilde{f}_i(m)\Phi(z_m(i))}$$

where T^- is the number of observations where $\mathbf{a}'\mathbf{O}_t \leq c$ and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and

$$z_m(i) = \frac{c - \mathbf{a}'\mu_{ic}(m)}{\sqrt{\mathbf{a}'\Sigma_{ic}(m)\mathbf{a}}}$$

In the case of no censoring, $C_{ic}(m) = 0$.

Viterbi algorithm

The major problem associated with fitting HSMMs is the estimation of the model parameters. Another problem is finding how best to assign a state to each observation once the parameters have been estimated, and to do this a variety of methods is available. The Viterbi algorithm (Forney 1973) finds the most likely sequence of states for the observations and does not permit any forbidden transitions.

The Viterbi algorithm requires the definition of two more probabilities. Firstly, $\delta_t(i)$ which is the maximum, over all possible state sequences p, of the probabilities of occurrence of the first t observations with the last one being the last of a series from S_i, i.e.,

$$\delta_t(i) = \max_p P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_1 = q_1, \ldots, Q_t = i, Q_{t+1} \neq i)$$

Secondly, $\delta_t^*(i)$ is the maximum, over p, of the probabilities of occurrence of the first t observations with a change to S_i starting at $t+1$, i.e.,

$$\delta_t^*(i) = \max_p P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_1 = q_1, \ldots, Q_t \neq i, Q_{t+1} = i)$$

It can be shown that

$$\begin{aligned} \delta_t(i) &= \max_d \delta_{t-d}^*(i) p_d(i) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s) \\ \delta_t^*(j) &= \max_i \delta_t(i) a_{ij} \end{aligned}$$

and note that $\delta_0^*(i) = \pi_i$. Both $\delta_t(i)$ and $\delta_t^*(j)$ become very small for sufficiently large T and $\log \delta_t(i)$ and $\log \delta_t^*(j)$ can be calculated instead. However, some of the factors involved can be zero, in which case the particular candidate can be ignored with regard to the maximum selection process before an error, due to taking the logarithm of zero, is generated.

It is also necessary to keep records of which d it was that maximised $\delta_t(i)$ and which i maximised $\delta_t^*(j)$ so that when the $\delta_T(i)$s have been determined it is possible to back-track through those records to recover the best sequence of states, q_t. Suppose these records are kept in $p_t(i)$ and $p_t^*(j)$ then

$$q_T = \operatorname*{argmax}_i \delta_T(i)$$

and

$$q_t = p_T^*(q_T) \quad t = T - p_T(q_T) + 1, \ldots, T$$

then the prior ones are found in a similar way with T replaced by $T - p_T(q_T)$ and so on until finally q_1 is determined.

The extent to which the Viterbi algorithm correctly attributes a state to an observation can be assessed through simulation. After estimating the model's parameters, simulated data can be generated and the Viterbi algorithm used on that data. Then the algorithms attributions can be compared to the simulated data where, of course, the state associated with each observation is known. Such an assessment is important since, for example, in one case it was found that, although the algorithm achieved a 95% correct attribution rate, one of the rarer states — 5% of the observations — was poorly recognised by the algorithm and only 1% of the data was attributed to that state.

References

Abramowitz, M. & Stegun, I. A. 1972: Handbook of mathematical functions. Dover Publications, New York.

Baum, L. E. & Petrie, T. 1966: Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics 37*: 1554–1563.

Billingsley, P. 1961: Statistical inference for Markov process. The University of Chicago Press, Chicago.

Cox, D. R. & Miller, H. D. 1965: The theory of stochastic processes. Methuen, London.

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977: Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society series B 39*: 1–38.

Elliot, R. J., Aggoun, L., & Moore, J. B. 1995: Hidden Markov models. Springer-Verlag, New York.

Ferguson, J. D. 1980: Variable duration models for speech. In Proceedings of the Symposium on Application of hidden Markov models to text and speech. Ferguson, J. D. (Ed.) Institute for Defense Analyses, Princeton NJ, 143–179.

Forney, G. D. 1973: The Viterbi algorithm. *Proceedings of the IEEE 61*: 268–278.

Johnson, N. L. & Kotz, S. 1972: Distributions in statistics. Wiley, New York

Juang 1984: On the hidden Markov model and dynamic time warping for speech recognition — A unified view. *AT&T Technical Journal 63*: 1213–1243.

Karlin, S. & Taylor, H. N. 1975: A first course in stochastic processes. Academic Press, New York

Levinson, S. E., Rabiner, L. R., & Sondhi, M. M. 1983: An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Systems Technical Journal 62*: 1035–1074.

MacDonald, I. L. & Zucchini, W. 1997: Hidden Markov and other models for discrete-valued time series. *Monographs on Statistics and Applied Probability No. 70.* Chapman & Hall, London.

Rabiner, L. R. 1989: A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*: 257–285.

Sansom, J. 1998: A hidden Markov model for rainfall using breakpoint data. *Journal of Climate 11*: 42–53.

Sansom, J. 1999: Large scale spatial variability of rainfall through hidden semi-Markov models of breakpoint data. *Journal of Geophysical Research 104 (D24)*: 31631–31643.

Sansom, J. & Thomson, P. J. 1998: Detecting components in censored and truncated meteorological data. *Envirometrics 9*: 673–688.

Zucchini, W. & Guttorp, P. 1991: A hidden Markov model for space-time precipitation. *Water Resources Research 27*: 1917–1923.

Appendix 1

Result 1 — Derivation of $\alpha_t(i)$

This is the probability of the occurrence of the first t observations with the last one being the last of a sequence from S_i i.e., define

$$\alpha_t(i) = P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t = i, Q_{t+1} \neq i)$$

thus

$$\begin{aligned}
\alpha_t(i) &= \sum_{d=1}^{t} P(\mathbf{O}_1, \ldots, \mathbf{O}_t, S_i \text{ begins at } t-d+1 \text{ and ends at } t) \\
&= P(\mathbf{O}_1, \ldots, \mathbf{O}_t, S_i \text{ begins at 1 and ends at } t) \\
&\quad + \sum_{d=1}^{t-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} P(\mathbf{O}_1, \ldots, \mathbf{O}_t, S_j \text{ ends at } t-d, S_i \text{ begins at } t-d+1 \text{ and ends at } t) \\
&= \pi_i p_t(i) \prod_{s=1}^{t} b_i(\mathbf{O}_s) + \sum_{j=1}^{N} \sum_{d=1}^{t-1} \alpha_{t-d}(j) a_{ji} p_d(i) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s)
\end{aligned}$$

This holds for $t = 1, \ldots, T$ provided the second term on the right is interpreted as zero when $t = 1$. Now defining, firstly

$$\alpha_t(i) = 0 \quad (t < 0)$$

and secondly

$$\alpha_0(i) = \pi_i \quad (i = 1, \ldots, N)$$

where π is the stationary distribution of the Markov chain given by $\pi \mathbf{A} = \pi$. Also it can be noted that

$$\sum_{j=1}^{N} \alpha_0(j) a_{ji} = \sum_{j=1}^{N} \pi_j a_{ji} = \pi_i$$

Thus for $t \geq 1$

$$\alpha_t(i) = \sum_{j=1}^{N} \sum_{d=1}^{\infty} \alpha_{t-d}(j) a_{ji} p_d(i) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s)$$

where the summation over d is now to ∞ since $\alpha_{t-d}(i) = 0$ for all $d > t$ and for $t = 1$

$$\begin{aligned}
\alpha_1(i) &= \sum_{j=1}^{N} \alpha_0(j) a_{ji} p_1(i) b_i(\mathbf{O}_1) \\
&= \pi_i p_1(i) b_i(\mathbf{O}_1)
\end{aligned}$$

from the second definition.

Results 2 and 3 — Derivation of $\alpha_t^*(i)$

This is the probability of the occurrence of the first t observations with a change to S_i starting at $t+1$ i.e.,

$$\alpha_t^*(i) = P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t \neq i, Q_{t+1} = i)$$

$$= \sum_{\substack{j=1 \\ j\neq i}}^{N} P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t = j, Q_{t+1} = i)$$

$$= \sum_{\substack{j=1 \\ j\neq i}}^{N} P(Q_{t+1} = i | Q_t = j, Q_{t+1} \neq j, \mathbf{O}_1, \ldots, \mathbf{O}_t) P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t = j, Q_{t+1} \neq j)$$

$$= \sum_{j=1}^{N} \alpha_t(j) a_{ji}$$

Substituting this into (5) gives

$$\alpha_t(i) = \sum_{d=1}^{\infty} \alpha^*_{t-d}(i) p_d(i) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s)$$

Also note that for $t = 0$ a value can be defined since from the definition for $\alpha_0(i)$ it follows that

$$\alpha^*_0(i) = \sum_{j=1}^{N} \alpha_0(j) a_{ji} = \alpha_0(i) = \pi_i$$

Result 4 — Derivation of $\beta_t(i)$

This is the probability of the occurrence of the observations from $t+1$ to T given that S_i ended at t i.e., define

$$\beta_t(i) = P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | Q_t = i, Q_{t+1} \neq i)$$

which, unlike the $\alpha_t(i)$, depends on T and is undefined for $t = T$ since otherwise $\mathbf{O}_{T+1}$ would be required. For $0 < t < T$

$$\beta_t(i) = \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{d=1}^{T-t} P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, S_j \text{ begins at } t+1 \text{ and ends at } t+d, Q_T \neq Q_{T+1} | Q_t = i, Q_{t+1} \neq i)$$

which by Bayes' theorem and using the substitution

$$P(S_j \text{ begins at } t+1 \text{ and ends at } t+d | Q_t = i, Q_{t+1} \neq i) = a_{ij}\, p_d(j)$$

can be written as

$$\beta_t(i) = \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{d=1}^{T-t} P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | Q_t = i, S_j \text{ begins at } t+1 \text{ and ends at } t+d) a_{ij}\, p_d(j)$$

$$= \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{d=1}^{T-t} a_{ij}\, P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_{t+d}) \times P(\mathbf{O}_{t+d+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | Q_{t+d} = j, Q_{t+d+1} \neq j) p_d(j)$$

$$= \sum_{j=1}^{N} a_{ij} \left(\sum_{d=1}^{T-t-1} \prod_{s=t+1}^{t+d} b_j(\mathbf{O}_s) P(\mathbf{O}_{t+d+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | Q_{t+d} = j, Q_{t+d+1} \neq j) p_d(j) + \prod_{s=t+1}^{T} b_j(\mathbf{O}_s) p_{T-t}(j) \right)$$

where the last term of the summation over d has been separated from the others as $t+d-1 = T+1$ when $d = T - t$ and no observation is available for O_{T+1}. However, by defining $\beta_T(i) = 1$ the

last term can be included with the others and by further defining $\beta_t(i) = 0$ for all $t > T$ then for $0 < t < T$

$$\beta_t(i) = \sum_{j=1}^{N} \sum_{d=1}^{\infty} a_{ij}\, \beta_{t+d}(j) p_d(j) \prod_{s=t+1}^{t+d} b_j(\mathbf{O}_s)$$

where the summation over d is now to ∞ since $\beta_{t+d}(i) = 0$ for all $d > T - t$.

Results 5 and 6 — Derivation of $\beta_t^*(i)$

This is the probability, given that S_i began at $t+1$, of the occurrence of the observations from $t+1$ to T i.e.,

$$\beta_t^*(i) = P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | Q_t \neq i, Q_{t+1} = i)$$

which is defined over the range $0 \leq t < T$. Now proceeding as for Result 4 by using Bayes theorem and noting that

$$P(S_i \text{ ends at } t+d | S_i \text{ begins at } t+1) = p_d(i)$$

gives

$$\begin{aligned}
\beta_t^*(i) &= \sum_{d=1}^{T-t} P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, S_i \text{ begins at } t+1 \text{ and} \\
&\qquad \text{ends at } t+d, Q_T \neq Q_{T+1} | Q_t \neq i, Q_{t+1} = i) \\
&= \sum_{d=1}^{T-t} P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | S_i \text{ begins at } t+1 \text{ and ends at } t+d) \\
&\qquad \times \; P(S_i \text{ ends at } t+d | S_i \text{ begins at } t+1) \\
&= \sum_{d=1}^{T-t-1} \textstyle\prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s) P(\mathbf{O}_{t+d+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | \\
&\qquad Q_{t+d} = i, Q_{t+d+1} \neq i) p_d(i) + \textstyle\prod_{s=t+1}^{T} b_i(\mathbf{O}_s) p_{T-t}(i)
\end{aligned}$$

and from the definitions for $\beta_t(i)$ where $t \geq T$ the summation over d can be taken to ∞ so

$$\beta_t^*(i) = \sum_{d=1}^{\infty} \beta_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s)$$

Substituting the right hand side of (10) in (9) gives

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, \beta_t^*(j)$$

which with (10) gives

$$\beta_t^*(i) = \sum_{j=1}^{N} \sum_{d=1}^{\infty} \beta_{t+d}^*(j) a_{ij}\, p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s)$$

Results 7 and 8 — Derivation of $\gamma_t(i)$

This is the probability that, given all the observations, the system was in S_i at time t i.e.,

$$\gamma_t(i) = P(Q_t = i | \mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1}) \quad (t = 1, \ldots, T)$$

$$= \sum_{d=1}^{T} P(Q_t = i \text{ and has duration } d | \mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1})$$

$$= \sum_{d=1}^{T} \sum_{\tau=\max(t-d,0)}^{\min(t-1,T-d)} P(S_i \text{ starts at } \tau+1 \text{ and ends at } \tau+d | \mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1})$$

$$= \sum_{d=1}^{T} \sum_{\tau=\max(t-d,0)}^{\min(t-1,T-d)} P(\mathbf{O}_1, \ldots, \mathbf{O}_T, S_i \text{ starts at } \tau+1 \text{ and ends at } \tau+d, Q_T \neq Q_{T+1})$$

$$\Big/ P(\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1})$$

However, when $0 < \tau < T - d$ the summand above can be re-written as the product of the following three probabilities

1) $P(\mathbf{O}_{\tau+d+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | \mathbf{O}_1, \ldots, \mathbf{O}_{\tau+d}, S_i \text{ starts at } \tau+1 \text{ and ends at } \tau+d)$
2) $P(\mathbf{O}_{\tau+1}, \ldots, \mathbf{O}_{\tau+d}, S_i \text{ starts at } \tau+1 \text{ and ends at } \tau+d | \mathbf{O}_1, \ldots, \mathbf{O}_\tau, Q_\tau \neq i, Q_{\tau+1} = i)$
3) $P(\mathbf{O}_1, \ldots, \mathbf{O}_\tau, Q_\tau \neq i, Q_{\tau+1} = i)$

which evaluates to

$$\beta_{\tau+d}(i) p_d(i) \prod_{s=\tau+1}^{\tau+d} b_i(\mathbf{O}_s) \alpha^*_\tau(i) \tag{19}$$

When $\tau = 0$ the summand becomes

$$\begin{aligned} &P(\mathbf{O}_1, \ldots, \mathbf{O}_T, S_i \text{ starts at 1 and ends at } d, Q_T \neq Q_{T+1}) \\ &\quad = P(\mathbf{O}_{d+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | \mathbf{O}_1, \ldots, \mathbf{O}_d, S_i \text{ starts at 1 and ends at } d) \\ &\qquad \times P(\mathbf{O}_1, \ldots, \mathbf{O}_d | S_i \text{ starts at 1 and ends at } d) p_d(i) \pi_i \\ &\quad = \beta_d(i) p_d(i) \prod_{s=1}^{d} b_i(\mathbf{O}_s) \alpha^*_0(i) \end{aligned}$$

which is the same as (19) when τ is set to zero. Similarly when $\tau = T - d$ the summand becomes

$$\begin{aligned} &P(\mathbf{O}_1, \ldots, \mathbf{O}_T, S_i \text{ starts at } T-d+1 \text{ and ends at } T) \\ &\quad = P(\mathbf{O}_{T-d+1}, \ldots, \mathbf{O}_T, S_i \text{ starts at } T-d+1 \text{ and ends at } T | \\ &\qquad \mathbf{O}_1, \ldots, \mathbf{O}_{T-d}, S_i \text{ starts at } T-d+1) \times P(\mathbf{O}_1, \ldots, \mathbf{O}_{T-d}, S_i \text{ starts at } T-d+1) \\ &\quad = \beta_T(i) p_d(i) \prod_{s=T-d+1}^{T} b_i(\mathbf{O}_s) \alpha^*_{T-d}(i) \end{aligned}$$

where $\beta_T(i)$ can be introduced as its value is 1. Thus (19) holds for

$$\max(t-d, 0) \leq \tau \leq \min(t-1, T-d)$$

and since $\alpha^*_\tau(i) = 0$ for $\tau < 0$ and $\beta_{\tau+d}(i) = 0$ for $\tau > T - d$ the limits for the summation over τ can be resolved as shown and

$$\gamma_t(i) = \sum_{d=1}^{T} \sum_{\tau=t-d}^{t-1} \beta_{\tau+d}(i) p_d(i) \prod_{s=\tau+1}^{\tau+d} b_i(\mathbf{O}_s) \alpha^*_\tau(i) \Big/ P(\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1})$$

Also from the definition for $\alpha_t(i)$ it can be noted that

$$P(\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1}) = \sum_{i=1}^{N} \alpha_T(i)$$

The expressions for $\alpha_t(i), \beta_i(i)$ *etc* are recursive and a recursive expression for $\gamma_t(i)$ can be found by considering

$$\begin{aligned}
\gamma_{t+1}(i) &= \sum_{d=1}^{T} \sum_{\tau=t+1-d}^{t} \beta_{\tau+d}(i) p_d(i) \prod_{s=\tau+1}^{\tau+d} b_i(\mathbf{O}_s) \alpha_\tau^*(i) \Big/ \sum_{i=1}^{N} \alpha_T(i) \\
&= \sum_{d=1}^{T} \left(\sum_{\tau=t-d}^{t-1} \beta_{\tau+d}(i) p_d(i) \prod_{s=\tau+1}^{\tau+d} b_i(\mathbf{O}_s) \alpha_\tau^*(i) + \beta_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s) \alpha_t^*(i) \right. \\
&\qquad \left. - \beta_t(i) p_d(i) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s) \alpha_{t-d}^*(i) \right) \Big/ \sum_{i=1}^{N} \alpha_T(i) \\
&= \gamma_t(i) + (\alpha_t^*(i)\beta_t^*(i) - \alpha_t(i)\beta_t(i)) \Big/ \sum_{i=1}^{N} \alpha_T(i) \qquad (20)
\end{aligned}$$

By substituting $t = 1$ in the full expression for $\gamma_t(i)$, noting that the only non-zero term in the summation over τ is for $\tau = 0$ and using (12) gives

$$\gamma_1(i) = \alpha_0^*(i)\beta_0^*(i) \Big/ \sum_{i=1}^{N} \alpha_T(i)$$

Successive substitutions in (20) yields for $t = 1, \ldots, T$

$$\gamma_t(i) = \left(\sum_{\tau=0}^{t-1} \alpha_\tau^*(i)\beta_\tau^*(i) - \sum_{\tau=1}^{t-1} \alpha_\tau(i)\beta_\tau(i) \right) \Big/ \sum_{i=1}^{N} \alpha_T(i)$$

where the second term in the denominator is interpreted as zero when $t = 1$.

Result 9 — Derivation of $\gamma_t(i,j)$

This is the probability that, given all the observations, the system was in S_i until time t and there was a change with S_j starting at time $t+1$ i.e., for $t = 1, \ldots, T-1$ and $i \neq j$

$$\begin{aligned}
\gamma_t(i,j) &= P(Q_t = i, Q_{t+1} = j | \mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1}) \\
&= P(\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_t = i, Q_{t+1} = j, Q_T \neq Q_{T+1}) / P(\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1}) \\
&= P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T, Q_T \neq Q_{T+1} | \mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t = i, Q_{t+1} = j) \\
&\quad \times P(Q_{t+1} = j | \mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t = i, Q_{t+1} \neq i) \\
&\quad \times P(\mathbf{O}_1, \ldots, \mathbf{O}_t, Q_t = i, Q_{t+1} \neq i) \Big/ \sum_{i=1}^{N} \alpha_T(i)
\end{aligned}$$

Thus for $t = 1, \ldots, T-1$ and $i \neq j$

$$\gamma_t(i,j) = \alpha_t(i) a_{ij} \beta_t^*(j) \Big/ \sum_{i=1}^{N} \alpha_T(i)$$

Although it is not generally required a value for when $i = j$ is given by

$$\gamma_t(i,i) = \gamma_t(i) - \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma_t(i,j)$$

Result 10 — Derivation of $\Delta_d(i)$

This is based on the probability that, given all the observations, the system was in S_i at time t and was in a visit of length d i.e., for $t = 1, \ldots, T$

$$\begin{aligned}
\delta_t(i,d) &= P(Q_t = i, \text{state in visit of duration d}|\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_{T+1} \neq Q_T) \\
&= \sum_{\tau=\max(t-d,0)}^{\min(t-1,T-d)} P(S_i \text{ starts at } \tau+1 \text{ and ends at } \tau+d|\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_{T+1} \neq Q_T) \\
&= \sum_{\tau=\max(t-d,0)}^{\min(t-1,T-d)} P(\mathbf{O}_1, \ldots, \mathbf{O}_T, S_i \text{ starts at } \tau+1 \text{ and ends at } \tau+d, Q_{T+1} \neq Q_T) \\
&\qquad \Big/ P(\mathbf{O}_1, \ldots, \mathbf{O}_T, Q_{T+1} \neq Q_T) \\
&= \sum_{\tau=t-d}^{t-1} \beta_{\tau+d}(i) p_d(i) \prod_{s=\tau+1}^{\tau+d} b_i(\mathbf{O}_s) \alpha_\tau^*(i) \Big/ \sum_{i=1}^{N} \alpha_T(i)
\end{aligned}$$

where the last step follows from the development leading to (19), but, rather than $\delta_t(i,d)$, it is $\Delta_d(i)$ that is required. Thus summing throughout over t and reversing the summation on the right hand side yields

$$\begin{aligned}
\Delta_d(i) &= \sum_{\tau=1-d}^{T-1} \sum_{t=\max(1,\tau+1)}^{\min(T,\tau+d)} \beta_{\tau+d}(i) p_d(i) \prod_{s=\tau+1}^{\tau+d} b_i(\mathbf{O}_s) \alpha_\tau^*(i) \Big/ \sum_{i=1}^{N} \alpha_T(i) \\
&= \sum_{\tau=0}^{T-1} \sum_{t=\tau+1}^{\min(T,\tau+d)} \beta_{\tau+d}(i) p_d(i) \prod_{s=\tau+1}^{\tau+d} b_i(\mathbf{O}_s) \alpha_\tau^*(i) \Big/ \sum_{i=1}^{N} \alpha_T(i)
\end{aligned}$$

since $\alpha_\tau^*(i) = 0$ for $\tau < 0$ and consequently since $\tau \geq 0$ thus $\tau + 1 \geq 1$. The inner summation's upper limit can always be taken to be $\tau + d$ as $\beta_{\tau+d}(i) = 0$ when $\tau + d > T$ and so it just yields d. Thus

$$\Delta_d(i) = d \sum_{t=0}^{T} \beta_{t+d}(i) p_d(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s) \alpha_t^*(i) \Big/ \sum_{i=1}^{N} \alpha_T(i)$$

Result 11

From Chapter 13 (p. 81) of Johnson & Kotz (1972) where X is a $\mathrm{N}(\mu, \sigma^2)$ random variable and with $A \leq X \leq B$ then, where $\phi(\cdot)$ is the standard normal probability density function and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function,

$$E\{X|A \leq X \leq B\} = \mu + \frac{\phi\left(\frac{A-\mu}{\sigma}\right) - \phi\left(\frac{B-\mu}{\sigma}\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)}\sigma$$

and

$$\mathrm{Var}(X|A \leq X \leq B) = \left[1 + \frac{\left(\frac{A-\mu}{\sigma}\right)\phi\left(\frac{A-\mu}{\sigma}\right) - \left(\frac{B-\mu}{\sigma}\right)\phi\left(\frac{B-\mu}{\sigma}\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)} - \left(\frac{\phi\left(\frac{A-\mu}{\sigma}\right) - \phi\left(\frac{B-\mu}{\sigma}\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)}\right)^2\right]\sigma^2$$

When concerned with values below a value c, then $A = -\infty$ and $B = c$ thus

$$\begin{aligned}\mu(c|\theta) &= \mu - \sigma\frac{\phi(\frac{c-\mu}{\sigma})}{\Phi(\frac{c-\mu}{\sigma})} \\ \sigma^2(c|\theta) &= \sigma^2\left(1 - \left(\frac{c-\mu}{\sigma}\right)\frac{\phi(\frac{c-\mu}{\sigma})}{\Phi(\frac{c-\mu}{\sigma})} - \left(\frac{\phi(\frac{c-\mu}{\sigma})}{\Phi(\frac{c-\mu}{\sigma})}\right)^2\right)\end{aligned}$$

Result 12

From Chapter 36 (p. 113) of Johnson & Kotz (1972) where $\mathbf{Z} = (Z_1, Z_2)'$ a $\text{MVN}_2(\mathbf{0}, \boldsymbol{\Sigma})$ random variable with $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = \rho$ and if $Z_1 > h, Z_2 > k$ then

$$\begin{aligned}E\{Z_1\} &= \Big(\phi(h)(1-\Phi(A)) + \rho\phi(k)(1-\Phi(B))\Big)\Big/L(h,k;\rho) \\ E\{Z_1^2\} &= \Big(h\phi(h)(1-\Phi(A)) + \rho^2 k\phi(k)(1-\Phi(B)) \\ &\qquad +\rho(1-\rho^2)^{\frac{1}{2}}\phi(h,k;\rho)\Big)\Big/L(h,k;\rho) + 1 \\ E\{Z_2\} &= \Big(\rho\phi(h)(1-\Phi(A)) + \phi(k)(1-\Phi(B))\Big)\Big/L(h,k;\rho) \\ E\{Z_2^2\} &= \Big(\rho^2 h\phi(h)(1-\Phi(A)) + k\phi(k)(1-\Phi(B)) \\ &\qquad +\rho(1-\rho^2)^{\frac{1}{2}}\phi(h,k;\rho)\Big)\Big/L(h,k;\rho) + 1 \\ E\{Z_1Z_2\} &= \Big(\rho h\phi(h)(1-\Phi(A)) + \rho k\phi(k)(1-\Phi(B)) \\ &\qquad +\rho(1-\rho^2)\phi(h,k;\rho)\Big)\Big/L(h,k;\rho) + \rho \qquad (21)\end{aligned}$$

where

$$\begin{aligned}A &= (k-\rho h^2)/\sqrt{1-\rho^2} \Rightarrow \Phi(A) \overset{k=-\infty}{\longrightarrow} 0 \\ B &= (h-\rho k^2)/\sqrt{1-\rho^2} \Rightarrow \Phi(B) \overset{k=-\infty}{\longrightarrow} 1 \\ \phi(h,k;\rho) &= \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(h^2 - 2\rho hk + k^2)\right) \overset{k=-\infty}{\longrightarrow} 0 \\ L(h,k;\rho) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \\ &\quad \times \int_h^\infty \int_k^\infty \exp\left(-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2)\right) dz_1 dz_2 \overset{k=\infty}{\longrightarrow} 0 \qquad (22)\end{aligned}$$

in which $k = -\infty$ represents no truncation of Z_2. Note that the bivariate form of Φ can be written as $\Phi(h,k;\rho)$ and $\Phi(h,\infty;\rho) = \Phi(h)$ also note that

$$\Phi(h,k;\rho) = 1 - L(h,-\infty;\rho) - L(-\infty,k;\rho) + L(h,k;\rho)$$

and so by letting $k = \infty$ when the last two terms are both zero

$$L(h,-\infty;\rho) = 1 - \Phi(h)$$

Substituting this and the results of (22) into (21) yields for truncation of Z_1 only

$$\begin{aligned}E\{Z_1\} &= \phi(h)/(1-\Phi(h)) \qquad (23) \\ \text{Var}(Z_1) &= 1 + h\frac{\phi(h)}{1-\Phi(h)} - \left(\frac{\phi(h)}{1-\Phi(h)}\right)^2 \qquad (24)\end{aligned}$$

$$
\begin{aligned}
E\{Z_2\} &= \rho E\{Z_1\} && (25)\\
\mathrm{Var}(Z_2) &= \rho^2 \mathrm{Var}(Z_1) - \rho^2 + 1 && (26)\\
E\{Z_1 Z_2\} &= \rho h E\{Z_1\} + \rho \\
\mathrm{Cov}(Z_1, Z_2) &= \rho \mathrm{Var}(Z_1) && (27)
\end{aligned}
$$

and so

$$
\begin{aligned}
E\{\mathbf{Z}|Z_1 > h\} &= \frac{\phi(h)}{1-\Phi(h)} \begin{pmatrix} 1 \\ \rho \end{pmatrix} \\
\mathrm{Var}(\mathbf{Z}|Z_1 > h) &= \mathrm{Var}(Z_1) \begin{pmatrix} 1 & \rho \\ \rho & \rho^2 + \frac{1-\rho^2}{\mathrm{Var}(Z_1)} \end{pmatrix}
\end{aligned}
$$

When concerned with $Z_1 < h$ and if, rather than standardised normal variates, X_1 is marginally distributed as $\mathrm{N}(\mu_1, \sigma_1^2)$ and X_2 as $\mathrm{N}(\mu_2, \sigma_2^2)$ then the following substitutions can be made: $Z_1 \to -\frac{X_1-\mu_1}{\sigma_1}$, $Z_2 \to -\frac{X_2-\mu_2}{\sigma_2}$ and $h \to -\frac{c-\mu_1}{\sigma_1}$. It should also be noted that $\phi(-c) = \phi(c)$ and $\Phi(-c) = 1 - \Phi(c)$. Thus from (23)

$$
E\{X_1\} = \mu_1 - \sigma_1 \frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})}
$$

and from (25)

$$
E\{X_2\} = \mu_2 - \sigma_2 \rho \frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})}
$$

and from (24)

$$
\mathrm{Var}(X_1) = \sigma_1^2 \left(1 - \left(\frac{c-\mu}{\sigma} \right) \frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})} - \left(\frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})} \right)^2 \right)
$$

and from (26)

$$
\mathrm{Var}(X_2) = \sigma_2^2 \left(1 - \rho^2 + \frac{\rho^2}{\sigma_1^2} \mathrm{Var}(X_1) \right)
$$

and finally from (27)

$$
\mathrm{Cov}(X_1, X_2) = \rho \frac{\sigma_2}{\sigma_1} \mathrm{Var}(X_1)
$$

and so

$$
\begin{aligned}
E\{\mathbf{X}|X_1 < h\} &= \begin{pmatrix} \mu_1 - \sigma_1 \frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})} \\ \mu_2 - \sigma_2 \rho \frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})} \end{pmatrix} \\
\mathrm{Var}(\mathbf{X}|X_1 < h) &= A \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2(\rho^2 + \frac{1}{A}(1-\rho^2)) \end{pmatrix} && (28)
\end{aligned}
$$

where

$$
A = \left(1 - \left(\frac{c-\mu_1}{\sigma_1} \right) \frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})} - \left(\frac{\phi(\frac{c-\mu_1}{\sigma_1})}{\Phi(\frac{c-\mu_1}{\sigma_1})} \right)^2 \right) \qquad (29)
$$

It can be noted from Abramowitz & Stegun (1972, p.932) that for $z < 0$

$$
\frac{\phi(z)}{\Phi(z)} = -z \Big/ \left(1 - \frac{1}{z^2} + \frac{1 \cdot 3}{z^4} - \frac{1 \cdot 3 \cdot 5}{z^6} + \cdots \right)
$$

which is most accurate for z large and negative but by including terms up to z^{14} values of A are in error by at most 0.1% for $z < -5$.

If truncation above an arbitrary line is required so that only the $\mathbf{X}$ that satisfy $\mathbf{a}'\mathbf{X} < c$ are available then this can be treated by a rotational transformation of the axes. If the (x_1, x_2) axes are rotated an angle ζ clockwise, then the co-ordinates of a point with respect to the new axes are given by

$$\begin{pmatrix} x_1^r \\ x_2^r \end{pmatrix} = \begin{pmatrix} \cos\zeta & \sin\zeta \\ -\sin\zeta & \cos\zeta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

or where $\mathbf{T}$ is the transformation matrix

$$\mathbf{X}^r = \mathbf{T}\mathbf{X}.$$

If $\mathbf{X}$ represents a $\mathrm{MVN}_2(\mu, \Sigma)$ variable then $\mathbf{X}^r$ will be a $\mathrm{MVN}_2(\mathbf{T}\mu, \mathbf{T}\Sigma\mathbf{T}')$ variable. Furthermore, suppose the arbitrary line is of slope g and passes through the point (p_1, p_2). In this case, the axes need to be rotated so that the line becomes vertical with the data to be retained on its left i.e., rotated an angle of $90 + \tan^{-1}(g) = 90 + \epsilon$, say so that ϵ is the angle between the line and the horizontal axes measured clockwise. Thus

$$\mathbf{T} = \begin{pmatrix} -\sin\epsilon & \cos\epsilon \\ -\cos\epsilon & -\sin\epsilon \end{pmatrix}$$

and $\mathbf{a}' = (-\sin\epsilon \ \cos\epsilon)$ i.e., the top row of $\mathbf{T}$ and since (p_1, p_2) is on the border of the truncation thus $c = p_2\cos\epsilon - p_1\sin\epsilon$. Let

$$\begin{aligned} \mathbf{T}\mu &= \begin{pmatrix} {}^r\mu_1 \\ {}^r\mu_2 \end{pmatrix} \\ \mathbf{T}\Sigma\mathbf{T}' &= \begin{pmatrix} {}^r\sigma_1^2 & {}^r\rho\,{}^r\sigma_1\,{}^r\sigma_2 \\ {}^r\rho\,{}^r\sigma_1\,{}^r\sigma_2 & {}^r\sigma_2^2 \end{pmatrix} \end{aligned}$$

where with a leading superscript, r, to denote that those parameters are for the rotated distribution, these expressions serve to define ${}^r\mu_1$, ${}^r\mu_2$, ${}^r\sigma_1$, ${}^r\sigma_2$ and ${}^r\rho$. After truncation the remaining data has mean μ_r and variance Σ_r and from (28) and (29)

$$\begin{aligned} \mu_r &= \mathbf{T}\mu - \frac{\phi(\frac{c-{}^r\mu_1}{{}^r\sigma_1})}{\Phi(\frac{c-{}^r\mu_1}{{}^r\sigma_1})} \begin{pmatrix} {}^r\sigma_1 \\ {}^r\rho\,{}^r\sigma_2 \end{pmatrix} \\ \Sigma_r &= {}^rA\mathbf{T}\Sigma\mathbf{T}' + {}^r\sigma_2^2(1-{}^r\rho^2)(1-{}^rA) \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

where (28) has been rewritten as

$$\mathrm{Var}(\mathbf{X}|X_1 < h) = {}^rA \begin{pmatrix} {}^r\sigma_1^2 & {}^r\rho\,{}^r\sigma_1\,{}^r\sigma_2 \\ {}^r\rho\,{}^r\sigma_1\,{}^r\sigma_2 & {}^r\sigma_2^2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & {}^r\sigma_2^2(1-{}^r\rho^2)(1-{}^rA) \end{pmatrix}$$

Now, noting that $\mathbf{T}^{-1} = \mathbf{T}'$, rotate back to obtain the mean $\mu_c = \mathbf{T}'\mu_r$ and variance $\Sigma_c = \mathbf{T}'\Sigma_r\mathbf{T}$ as

$$\begin{aligned} \mu_c &= \mu - \frac{\phi(\frac{c-{}^r\mu_1}{{}^r\sigma_1})}{\Phi(\frac{c-{}^r\mu_1}{{}^r\sigma_1})}\mathbf{T}' \begin{pmatrix} {}^r\sigma_1 \\ {}^r\rho\,{}^r\sigma_2 \end{pmatrix} \\ \Sigma_c &= {}^rA\Sigma + {}^r\sigma_2^2(1-{}^r\rho^2)(1-{}^rA) \begin{pmatrix} \cos^2\epsilon & \sin\epsilon\cos\epsilon \\ \sin\epsilon\cos\epsilon & \sin^2\epsilon \end{pmatrix} \end{aligned}$$

These are the mean and variance of the data that satisfy $\mathbf{a}'\mathbf{X} < c$.