Mathematical Background Notes for Package "HiddenMarkov"

David Harte
Statistics Research Associates
PO Box 12 649
Wellington 6144
NEW ZEALAND
Email: david@statsresearch.co.nz

27 November 2019

Abstract

These notes give a very brief background to some relationships that are used in the R package "HiddenMarkov" (Harte, 2010). This package fits various hidden Markov models. R is a comprehensive statistical programming language managed by the R Development Core Team (2010).

Contents

1 Discrete Time Hidden Markov Model

1.1 Markov Chain

$\{C_i; i = 1, \cdots, n\}$ has m states $\{1, \cdots, m\}$. It satisfies the *Markov Property*:

$$\begin{aligned}\Pr\{C_i \,|\, C_{i-1}, \cdots, C_1\} &= \Pr\{C_i \,|\, C_{i-1}\} \\ &= \Pr\{C_i = k \,|\, C_{i-1} = j\} \\ &= \gamma_{jk}^{(i)} .\end{aligned}$$

If $\gamma_{jk}^{(i)} = \gamma_{jk}, \ \forall i$ and $j, k = 1, \cdots, m$, then $\{C_i\}$ is *homogeneous*.

Note: we use the subscript i to denote the discrete time points, and j and k to denote the Markov states.

Now assume that $\{C_i\}$ is homogeneous. Let $\Gamma = (\gamma_{jk})$ be an $m \times m$ transition matrix. Let $\delta_j^{(i)} = \Pr\{C_i = j\}$, and

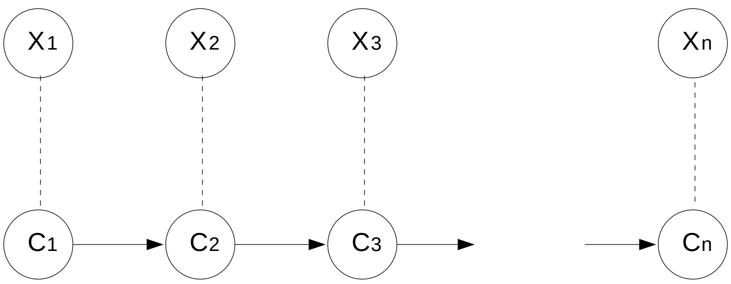
$$\delta^{(i)} = (\delta_1^{(i)}, \delta_2^{(i)}, \cdots, \delta_m^{(i)}) ,$$

then

$$\delta^{(i)} = \delta^{(i-1)}\Gamma = \delta^{(i-2)}\Gamma^2 = \delta^{(i-3)}\Gamma^3 .$$

The chain is *stationary* if $\delta^{(i)} = \delta \quad \forall i$, i.e. $\delta = \delta\Gamma$.

1.2 The Model



Denote the history of the process until time i as $X^{(i)}$.

Has *conditional independence*

$$\Pr\{X_i \,|\, X^{(i-1)}, C^{(i)}\} = \Pr\{X_i \,|\, C_i\} .$$

When X_i is a continuous random variable, replace the probability function with the density function.

Let

$$p_{ij} = \Pr\{X_i = x_i \,|\, C_i = j\}\,,$$

and

$$D_i = \text{diag}(p_{i1}, p_{i2}, \cdots, p_{im})\,.$$

Further, let Λ be the set of parameters relevant to the observed probability distribution p_{ij}. We denote the set of model parameters $(\delta, \Gamma, \Lambda)$ collectively as Θ.

1.3 Forward and Backward Probabilities

The *forward* probabilities are

$$\alpha_{ij} = \Pr\{X_1 = x_1, \cdots, X_i = x_i, C_i = j\}$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, m$. They are calculated in a "forward" recursive manner. So

$$\alpha_{1j} = \Pr\{X_1 = x_1, C_1 = j\} = \Pr\{X_1 = x_1 \,|\, C_1 = j\} \Pr\{C_1 = j\} = \delta_j^{(1)} p_{1j}$$

then

$$\begin{aligned}
\alpha_{2j} &= \Pr\{X_1 = x_1, X_2 = x_2, C_2 = j\} \\
&= \sum_{k=1}^m \Pr\{X_1 = x_1, X_2 = x_2, C_1 = k, C_2 = j\} \\
&= \sum_{k=1}^m \Pr\{X_1 = x_1, X_2 = x_2 \,|\, C_1 = k, C_2 = j\} \Pr\{C_1 = k, C_2 = j\} \\
&= \sum_{k=1}^m \Pr\{X_1 = x_1 \,|\, C_1 = k\} \Pr\{X_2 = x_2 \,|\, C_2 = j\} \Pr\{C_2 = j \,|\, C_1 = k\} \Pr\{C_1 = k\} \\
&= \sum_{k=1}^m \alpha_{1k} \gamma_{kj} p_{2j} \\
&= \sum_{k=1}^m \delta_k^{(1)} p_{1k} \gamma_{kj} p_{2j}\,,
\end{aligned}$$

and so

$$(\alpha_{21}, \cdots, \alpha_{2m}) = \delta^{(1)} D_1 \Gamma D_2\,.$$

Similarly, it can be shown that

$$(\alpha_{i1}, \cdots, \alpha_{im}) = \delta^{(1)} D_1 (\Gamma D_2) \cdots (\Gamma D_i)\,. \tag{1}$$

The *backward* probabilities are

$$\beta_{ij} = \Pr\{X_{i+1} = x_{i+1}, \cdots, X_n = x_n \,|\, C_i = j\}$$

for $i = 1, \cdots, n-1$ and $j = 1, \cdots, m$. They are calculated in a "backward" recursive manner. Initially we set

$$(\beta_{n1}, \cdots, \beta_{nm}) = (1, \cdots, 1)_{1\times m}.$$

Then

$$\begin{aligned}
\beta_{(n-1)j} &= \Pr\{X_n = x_n \,|\, C_{n-1} = j\} \\
&= \Pr\{X_n = x_n, C_{n-1} = j\} / \Pr\{C_{n-1} = j\} \\
&= \sum_{k=1}^{m} \Pr\{X_n = x_n, C_{n-1} = j, C_n = k\} / \Pr\{C_{n-1} = j\} \\
&= \sum_{k=1}^{m} \Pr\{X_n = x_n \,|\, C_{n-1} = j, C_n = k\} \Pr\{C_{n-1} = j, C_n = k\} / \Pr\{C_{n-1} = j\} \\
&= \sum_{k=1}^{m} \Pr\{X_n = x_n \,|\, C_n = k\} \Pr\{C_n = k \,|\, C_{n-1} = j\} \,,
\end{aligned}$$

and so

$$(\beta_{(n-1)1}, \cdots, \beta_{(n-1)m})' = \Gamma D_n 1' \,.$$

Similarly,

$$(\beta_{i1}, \cdots, \beta_{im})' = (\Gamma D_{i+1})(\Gamma D_{i+2}) \cdots (\Gamma D_n) 1' \,. \tag{2}$$

Given estimates of the model parameters Θ, the $n \times m$ matrices $A = (\alpha_{ij})$ and $B = (\beta_{ij})$ can be calculated in a recursive manner.

1.4 Likelihood Function

Let $1' = (1, \cdots, 1)_{1 \times m}$. Note that

$$\begin{aligned}
\Pr\{X_i = x_i\} &= \sum_{j=1}^{m} \Pr\{X_i = x_i \,|\, C_i = j\} \Pr\{C_i = j\} \\
&= \delta^{(i)} D_i 1' \,,
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\{X_i = x_i, X_{i+1} = x_{i+1}\} \\
&\quad = \sum_{k_i=1}^{m} \sum_{k_{i+1}=1}^{m} \Pr\{X_i = x_i, X_{i+1} = x_{i+1} \,|\, C_i = k_i, C_{i+1} = k_{i+1}\} \Pr\{C_i = k_i, C_{i+1} = k_{i+1}\} \\
&\quad = \sum_{k_i=1}^{m} \sum_{k_{i+1}=1}^{m} \Pr\{X_i = x_i \,|\, C_i = k_i\} \Pr\{X_{i+1} = x_{i+1} \,|\, C_{i+1} = k_{i+1}\} \Pr\{C_i = k_i\} \\
&\qquad\qquad \Pr\{C_{i+1} = k_{i+1} \,|\, C_i = k_i\} \\
&\quad = \delta^{(i)} D_i \Gamma D_{i+1} 1' \,,
\end{aligned}$$

and also

$$\Pr\{X_i = x_i, X_{i+\ell} = x_{i+\ell}\} = \delta^{(i)} D_i \Gamma^{\ell} D_{i+\ell} 1' \,.$$

Similarly

$$\begin{aligned}
L = \Pr\{X^{(n)} = x^{(n)}\} &= \Pr\{X_1 = x_1, \cdots, X_n = x_n\} \\
&= \delta^{(1)} D_1 \Gamma D_2 \Gamma D_3 \cdots \Gamma D_n 1' \\
&= \delta^{(1)} D_1 (\Gamma D_2)(\Gamma D_3) \cdots (\Gamma D_n) 1' \,.
\end{aligned}$$

If stationary, $\delta^{(1)}$ can be replaced with $\delta = \delta\Gamma$, creating a recursive pattern ΓD_i for $i = 1, \cdots, n$.

Note the relationship with the *forward* (Eq 1) and *backward* (Eq 2) probabilities, i.e. for $\forall i = 1, \cdots, n$,

$$
\begin{aligned}
L &= \delta^{(1)} D_1 (\Gamma D_2) \cdots (\Gamma D_i)(\Gamma D_{i+1})(\Gamma D_{i+2}) \cdots (\Gamma D_n) 1' \\
&= (\alpha_{i1}, \cdots, \alpha_{im})(\beta_{i1}, \cdots, \beta_{im})' \,. \qquad (3)
\end{aligned}
$$

We want to estimate all parameters in $\Theta = (\delta, \Gamma, \Lambda)$ by maximising L. To do this, we consider the *complete data likelihood*.

1.5 Complete Data Likelihood

$$
\begin{aligned}
L_c &= \Pr\{X_1 = x_1, \cdots, X_n = x_n, C_1 = c_1, \cdots, C_n = c_n\} \\
&= \Pr\{X_1 = x_1, \cdots, X_n = x_n \,|\, C_1 = c_1, \cdots, C_n = c_n\} \\
&\qquad \Pr\{C_1 = c_1, \cdots, C_n = c_n\} \\
&= \Pr\{X_1 = x_1 \,|\, C_1 = c_1\} \Pr\{C_1 = c_1\} \\
&\qquad \prod_{i=2}^{n} \Pr\{X_i = x_i \,|\, C_i = c_i\} \Pr\{C_i = c_i \,|\, C_{i-1} = c_{i-1}\} \\
&= \delta_{c_1}^{(1)} \gamma_{c_1 c_2} \gamma_{c_2 c_3} \cdots \gamma_{c_{n-1} c_n} \prod_{i=1}^{n} \Pr\{X_i = x_i \,|\, C_i = c_i\}
\end{aligned}
$$

Now let

$$
\begin{aligned}
p_{ij} &= \Pr\{X_i = x_i \,|\, C_i = j\} \\
u_{ij} &= \begin{cases} 1 & \text{if } C_i = j \\ 0 & \text{otherwise} \end{cases} \\
v_{ijk} &= \begin{cases} 1 & \text{if } C_{i-1} = j \text{ and } C_i = k \\ 0 & \text{otherwise} \,. \end{cases}
\end{aligned}
$$

Then

$$
\log L_c = \sum_{j=1}^{m} u_{1j} \log \delta_j^{(1)} + \sum_{j=1}^{m} \sum_{k=1}^{m} \left(\sum_{i=2}^{n} v_{ijk} \right) \log \gamma_{jk} + \sum_{j=1}^{m} \sum_{i=1}^{n} u_{ij} \log p_{ij} \,.
$$

1.6 Baum-Welch Algorithm (EM)

Recall that

$$
L_c = \Pr\{X^{(n)} = x^{(n)}, C^{(n)} = c^{(n)}\} \,,
$$

so that

$$\underbrace{L_c}_{\text{maximise in M-step}} = \underbrace{\Pr\{C^{(n)} = c^{(n)} \,|\, X^{(n)} = x^{(n)}\}}_{\text{calculate in E-step}} \Pr\{X^{(n)} = x^{(n)}\}\,.$$

1.6.1 Outline of Procedure

1. Guess initial values for $\widehat{\Theta}$.

2. Start Loop.

3. *E-Step*: estimate u_{ij} and v_{ijk} given $\widehat{\Theta}$ (i.e. current estimate of Θ), by taking their conditional *expectations*, i.e.:

$$\begin{aligned} \widehat{u}_{ij} &= \mathrm{E}[u_{ij} \,|\, \widehat{\Theta}] \\ &= \Pr\{C_i = j \,|\, X^{(n)} = x^{(n)}, \widehat{\Theta}\} \\ &= \widehat{\alpha}_{ij}\widehat{\beta}_{ij}/\widehat{L} \end{aligned}$$

   and

$$\begin{aligned} \widehat{v}_{ijk} &= \mathrm{E}[v_{ijk} \,|\, \widehat{\Theta}] \\ &= \Pr\{C_{i-1} = j, C_i = k \,|\, X^{(n)} = x^{(n)}, \widehat{\Theta}\} \\ &= \widehat{\gamma}_{jk}\widehat{\alpha}_{i-1,j}\widehat{p}_{ik}\widehat{\beta}_{ik}/\widehat{L}\,. \end{aligned}$$

4. *M-Step*: estimate new values for $\widehat{\Theta}$ by *maximising* L_c; see §1.6.2, §1.6.3, and §1.6.4.

5. If $\widehat{\Theta}$ not converged, return to (2).

6. Stop.

If the Markov chain is non-stationary, the *M-step* can be performed by maximising each term in L_c separately.

1.6.2 First Term of L_c

Want to maximise

$$\sum_{j=1}^{m} u_{1j} \log \delta_j^{(1)}$$

subject to

$$\sum_{j=1}^{m} \delta_j^{(1)} = 1\,.$$

Let

$$F = \sum_{j=1}^{m} u_{1j} \log \delta_j^{(1)} + \theta \left(1 - \sum_{j=1}^{m} \delta_j^{(1)}\right)$$

where θ is a Lagrange multiplier. Then

$$\frac{\partial F}{\partial \delta_j^{(1)}} = \frac{u_{1j}}{\delta_j^{(1)}} - \theta$$

so that $\theta = u_{1j}/\delta_j^{(1)}$ for all j, hence

$$\widehat{\delta}_j^{(1)} = u_{1j}\,.$$

1.6.3 Second Term of L_c

Similarly as above, let

$$F = \sum_{j=1}^m \sum_{k=1}^m \left(\sum_{i=2}^n v_{ijk} \right) \log \gamma_{jk} + \sum_{j=1}^m \theta_j \left(1 - \sum_{k=1}^m \gamma_{jk} \right)$$

where $\theta_1, \cdots, \theta_m$ are Lagrange multipliers. Thus

$$\frac{\partial F}{\partial \gamma_{jk}} = -\theta_j + \frac{1}{\gamma_{jk}} \sum_{i=2}^n v_{ijk}\,,$$

hence letting $-\theta_j \gamma_{jk} + \sum_{i=2}^n v_{ijk} = 0$, we get

$$\sum_{k=1}^m \left(-\theta_j \gamma_{jk} + \sum_{i=2}^n v_{ijk} \right) = 0\,.$$

Since $\sum_{k=1}^m \gamma_{jk} = 1$, then

$$\theta_j = \sum_{k=1}^m \sum_{i=2}^n v_{ijk}\,,$$

so that

$$\widehat{\gamma}_{jk} = \frac{\sum_{i=2}^n v_{ijk}}{\sum_{k=1}^m \sum_{i=2}^n v_{ijk}}\,.$$

1.6.4 Third Term of L_c

Maximisation of the last term, i.e.

$$\sum_{j=1}^m \sum_{i=1}^n u_{ij} \log p_{ij}$$

depends on the probability distribution of the observed process, i.e. $p_{ij} = \Pr\{X_i = x_i \,|\, C_i = j\}$. The set of parameters is denoted by Λ.

The following subsections give details for specific distributions.

Poisson Distribution

In this case

$$p_{ij} = \Pr\{X_i = x_i \,|\, C_i = j\} = \frac{\lambda_j^{x_i}}{x_i!} \exp(-\lambda_j) \,.$$

Let

$$\begin{aligned} F &= \sum_{j=1}^m \sum_{i=1}^n u_{ij} \log p_{ij} \\ &= \sum_{j=1}^m \sum_{i=1}^n u_{ij} \left[x_i \lambda_j - \log(x_i!) - \lambda_j \right] \,, \end{aligned}$$

and so

$$\frac{\partial F}{\partial \lambda_j} = \frac{1}{\lambda_j} \sum_{i=1}^n u_{ij}(x_i - 1) \,,$$

hence

$$\widehat{\lambda}_j = \frac{\sum_{i=1}^n u_{ij} x_i}{\sum_{i=1}^n u_{ij}} \,.$$

Exponential Distribution

In this case

$$p_{ij} = f_{X_i}(x_i \,|\, C_i = j) = \lambda_j \exp(-\lambda_j x_i) \,.$$

Let

$$\begin{aligned} F &= \sum_{j=1}^m \sum_{i=1}^n u_{ij} \log p_{ij} \\ &= \sum_{j=1}^m \sum_{i=1}^n u_{ij} \left[\log \lambda_j - \lambda_j x_i \right] \,, \end{aligned}$$

and so

$$\frac{\partial F}{\partial \lambda_j} = \sum_{i=1}^n u_{ij} \left(\frac{1}{\lambda_j} - x_i \right) \,,$$

hence

$$\widehat{\lambda}_j = \frac{\sum_{i=1}^n u_{ij}}{\sum_{i=1}^n u_{ij} x_i} \,.$$

Binomial Distribution

In this case

$$p_{ij} = \Pr\{X_i = x_i \,|\, C_i = j\} = \begin{pmatrix} n_i \\ x_i \end{pmatrix} \pi_j^{x_i} (1 - \pi_j)^{n_i - x_i} \,,$$

and so

$$\widehat{\pi}_j = \frac{\sum_{i=1}^n u_{ij} x_i}{\sum_{i=1}^n u_{ij} n_i} \,.$$

Gaussian Distribution

In this case

$$p_{ij} = f_{X_i}(x_i \,|\, C_i = j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-1}{2\sigma_j^2}(x_i - \mu_j)^2 \right) ,$$

and so

$$\widehat{\mu}_j = \frac{\sum_{i=1}^n u_{ij} x_i}{\sum_{i=1}^n u_{ij}}$$

and

$$\widehat{\sigma}_j = \sqrt{\frac{\sum_{i=1}^n u_{ij}(x_i - \widehat{\mu}_j)^2}{\sum_{i=1}^n u_{ij}}} .$$

Gamma Distribution

$$f(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} \exp(-\lambda x)$$

$$\begin{aligned} F &= \frac{1}{n} \sum_{i=1}^n \log f(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n [a \log \lambda - \log \Gamma(a) + (a-1) \log x_i - \lambda x_i] \\ &= a \log \lambda - \log \Gamma(a) + (a-1) \overline{\log x} - \lambda \overline{x} \end{aligned}$$

$$\begin{aligned} \frac{\partial F}{\partial \lambda} &= \frac{a}{\lambda} - \overline{x} \\ \frac{\partial F}{\partial a} &= \log \lambda - \Psi(a) + \overline{\log x} \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 F}{\partial \lambda^2} &= \frac{-a}{\lambda^2} \\ \frac{\partial^2 F}{\partial a^2} &= -\Psi'(a) \\ \frac{\partial^2 F}{\partial a \partial \lambda} &= \frac{\partial^2 F}{\partial \lambda \partial a} = \frac{1}{\lambda} \end{aligned}$$

$$\begin{pmatrix} \lambda' \\ a' \end{pmatrix} = \begin{pmatrix} \lambda \\ a \end{pmatrix} - \begin{pmatrix} \frac{-a}{\lambda^2} & \frac{1}{\lambda} \\ \frac{1}{\lambda} & -\Psi'(a) \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial F}{\partial \lambda} \\ \frac{\partial F}{\partial a} \end{pmatrix}$$

The two sufficient statistics $\overline{x}$ and $\overline{\log x}$ become, for $j = 1, \cdots, m$,

$$\frac{\sum_{i=1}^n u_{ij} x_i}{\sum_{i=1}^n u_{ij}} \quad \text{and} \quad \frac{\sum_{i=1}^n u_{ij} \log x_i}{\sum_{i=1}^n u_{ij}} .$$

Beta Distribution

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

$$\begin{aligned} F &= \frac{1}{n}\sum_{i=1}^{n} \log f(x_i) \\ &= \log\Gamma(a+b) - \log\Gamma(a) - \log\Gamma(b) + (a-1)\overline{\log x} \\ &\qquad +(b-1)\overline{\log(1-x)} \end{aligned}$$

$$\begin{aligned} \frac{\partial F}{\partial a} &= \Psi(a+b) - \Psi(a) + \overline{\log x} \\ \frac{\partial F}{\partial b} &= \Psi(a+b) - \Psi(a) + \overline{\log(1-x)} \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 F}{\partial a^2} &= \Psi'(a+b) - \Psi'(a) \\ \frac{\partial^2 F}{\partial b^2} &= \Psi'(a+b) - \Psi'(b) \\ \frac{\partial^2 F}{\partial a \partial b} &= \frac{\partial^2 F}{\partial b \partial a} = \Psi'(a+b) \end{aligned}$$

$$\begin{pmatrix} a' \\ b' \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} - \begin{pmatrix} \Psi'(a+b) - \Psi'(a) & \Psi'(a+b) \\ \Psi'(a+b) & \Psi'(a+b) - \Psi'(b) \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial F}{\partial a} \\ \frac{\partial F}{\partial b} \end{pmatrix}$$

The two sufficient statistics $\overline{\log x}$ and $\overline{\log(1-x)}$ become, for $j = 1, \cdots, m$,

$$\frac{\sum_{i=1}^n u_{ij} \log x_i}{\sum_{i=1}^n u_{ij}} \quad \text{and} \quad \frac{\sum_{i=1}^n u_{ij} \log(1-x_i)}{\sum_{i=1}^n u_{ij}} .$$

Log Normal Distribution

If X has a lognormal distribution with parameters μ and σ, then $\log X$ has a normal distribution with mean μ and variance σ^2. In this case

$$p_{ij} = f_{X_i}(x_i \,|\, C_i = j) = \frac{1}{\sqrt{2\pi}\sigma_j x_i} \exp\left(\frac{-1}{2\sigma_j^2} (\log x_i - \mu_j)^2 \right) ,$$

and so

$$\begin{aligned}\text{E}[\log X_i \,|\, C_i = j] &= \mu_j\,, \\ \text{Var}[\log X_i \,|\, C_i = j] &= \sigma_j^2\,, \\ \text{E}[X_i \,|\, C_i = j] &= \exp(\mu_j + \sigma_j^2/2)\,, \text{ and} \\ \text{Var}[X_i \,|\, C_i = j] &= \exp(2\mu_j + \sigma_j^2)(\exp(\sigma_j^2) - 1)\,.\end{aligned}$$

Further

$$\widehat{\mu}_j = \frac{\sum_{i=1}^n u_{ij} \log x_i}{\sum_{i=1}^n u_{ij}}$$

and

$$\widehat{\sigma}_j = \sqrt{\frac{\sum_{i=1}^n u_{ij}(\log x_i - \widehat{\mu}_j)^2}{\sum_{i=1}^n u_{ij}}}\,.$$

Logistic Distribution

Like the beta and gamma distributions, a Newton iterative procedure is used here too. The required first and second derivatives can be found in Rao & Hamed (2000, §9.1.2). Here the location parameter is denoted by m and the scale parameter by a. Note that there are a couple of errors:

In Equation 9.1.10, n should be N; and Equation 9.1.11 should be

$$y_i = 1 + \exp\left(\frac{-(x_i - m)}{a}\right)\,.$$

Equation 9.1.19 should be

$$\frac{\partial^2}{\partial m^2} \log L = \frac{2}{a^2} \sum_{i-1}^{N} (y_i^{-2} - y_i^{-1})\,.$$

1.7 Pseudo Residuals

We follow the method outlined by Zucchini (2005). Let $X^{(-i)} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$, i.e. denotes the observed process except for the point X_i. For each $i = 1, \cdots, n$ we calculate

$$\begin{aligned}\psi_i &= \Pr\{X_i \leq x_i \,|\, X^{(-i)} = x^{(-i)}\} \\ &= \frac{\Pr\{X_i \leq x_i, X^{(-i)} = x^{(-i)}\}}{\Pr\{X^{(-i)} = x^{(-i)}\}} \\ &= \frac{\delta^{(1)} D_1 (\Gamma D_2) \cdots (\Gamma D_{i-1})(\Gamma D_i')(\Gamma D_{i+1})(\Gamma D_{i+2}) \cdots (\Gamma D_n) 1'}{\delta^{(1)} D_1 (\Gamma D_2) \cdots (\Gamma D_{i-1})(\Gamma I)(\Gamma D_{i+1})(\Gamma D_{i+2}) \cdots (\Gamma D_n) 1'}\,,\end{aligned}$$

where D_i' is an $m \times m$ diagonal matrix with elements $\Pr\{X_i \leq x_i \,|\, C_i = j\}$ for $j = 1, \cdots, m$, and I is the identity matrix. This is achieved by using the forward and backward probabilities.

The pseudo residuals are then $z_i = \Phi^{-1}(\psi_i)$, where Φ denotes the standard normal distribution function. If the observation sequence has been sampled from the assumed model, then the z_i's should have an approximate standard normal distribution.

If the distribution of the observation variables is discrete the following correction is made. Also calculate $\psi_i' = \Pr\{X_i \le x_i - 1 \,|\, X^{(-i)} = x^{(-i)}\}$, then

$$z_i = \Phi^{-1}\left(\frac{\psi_i + \psi_i'}{2}\right) .$$

1.8 Viterbi Algorithm

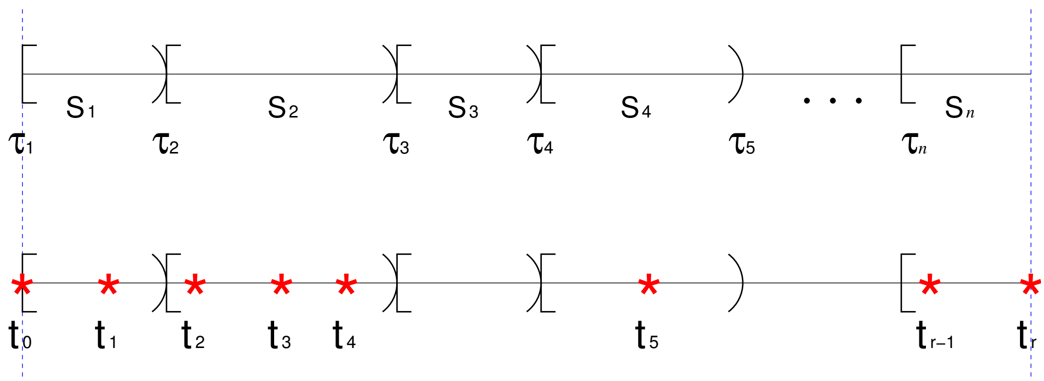
The purpose of the Viterbi algorithm is to *globally decode* the underlying hidden Markov state at each time point. It does this by determining the sequence of states $(c_1^*, \cdots, c_n^*)$ which maximises the joint distribution of the hidden states given the entire observed process $\{x^{(n)}\}$, i.e.

$$(c_1^*, \cdots, c_n^*) = \underset{c_1,\cdots,c_n \in \{1,2,\cdots,m\}}{\text{argmax}} \Pr\{C_1 = c_1, \cdots, C_n = c_n \,|\, X^{(n)} = x^{(n)}\} .$$

2 Markov Modulated Poisson Process

2.1 The Model

Let $S(t)$ be a Markov process in continuous time having discrete states $1, \cdots, m$. The process makes a transition from state s_{i-1} to s_i at time τ_i. The time spent in state s_i has an exponential distribution with parameter q_{s_i}. Events are occur as a Poisson process at times $t_1, t_2, \cdots, t_r$. The Poisson rate is determined by the Markov state, being constant within each Markov state.



We use the same formulation of the model as Rydén (1996). He assumes that the start and finish of the observation period coincides with events at times t_0 and t_r. See also Meier-Hellstern (1987) for further discussion.

2.2 Q Matrix

Let $P(t)$ be an $m \times m$ matrix with elements

$$p_{jk}(t) = \Pr\{S(t) = k \mid S(0) = j\} \,,$$

where $S(t)$ is a continuous time Markov process with m discrete states.

Let Q be the $m \times m$ *infinitesimal generator matrix* with jkth element q_{jk}, such that

$$\frac{d}{dt} p_{jk}(t) = \sum_\ell p_{j\ell}(t) q_{\ell k} = \sum_\ell q_{j\ell} p_{\ell k}(t) \,.$$

Given the initial condition that $P(0) = I$, $P(t)$ has solution

$$P(t) = \exp(tQ) \,.$$

Note that the diagonal elements are negative, and $-q_{jj}$ and is the exponential rate of transitions out of state j. Letting $q_j = -q_{jj}$, q_{jk}/q_j are the transition probabilities from state j to state k when $j \neq k$. Hence

$$\sum_{k=1}^m q_{jk} = 0$$

for all j.

2.3 Matrix Exponential

Given an $m \times m$ matrix Q, $\exp(Q)$ is to be interpreted as the *matrix exponential*, not the exponential of the individual elements. We briefly outline various methods of evaluation.

2.3.1 Taylor's Series Expansion

$$\exp(Q) = I + Q + \frac{1}{2!} Q^2 + \frac{1}{3!} Q^3 + \cdots$$

where I is the $m \times m$ identity matrix.

2.3.2 Eigen Value Decomposition

Assume that there exists a matrix E of eigenvectors and an $m \times m$ diagonal matrix Ψ containing the eigenvalues $\psi_1, \cdots, \psi_m$ such that $\Lambda - Q = E \Psi E^{-1}$, where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_m)$. Note that if Ψ is a diagonal matrix, then $\exp(\Psi)$ is also a diagonal matrix. Inserting this into a Taylor's series expansion gives

$$\exp(\Lambda - Q) = I + E\Psi E^{-1} + \frac{1}{2!} E \Psi^2 E^{-1} + \frac{1}{3!} E \Psi^3 E^{-1} + \cdots = E \exp(\Psi) E^{-1} \,.$$

2.3.3 Poisson Series Expansion

Let a be a number that is a little larger than the absolute values of the elements on the diagonal of the $m \times m$ matrix Q; and define B as

$$B = I + \frac{1}{a}Q\,,$$

where I is the $m \times m$ identity matrix. Then $Q = a(B - I)$, and so

$$\begin{aligned}\exp(Q) &= \exp(a(B-I)) \\ &= \exp(aB)\exp(-a) \\ &= I\exp(-a) + B\frac{a^1}{1!}\exp(-a) + B^2\frac{a^2}{2!}\exp(-a) + B^3\frac{a^3}{3!}\exp(-a) + \cdots\,.\end{aligned}$$

See Klemm et al (2003, §2.2) for further details.

2.4 Likelihood Function

2.4.1 Conditional Independence and Likelihood

Assume that events occur at times $t_0, t_1, \cdots, t_r$, that $t_0 = 0$, and that t_r coincides with the end of the observation period. Let $y_\ell = t_\ell - t_{\ell-1}$ for $\ell = 1, \cdots, r$. Also define the *auxiliary Markov chain* as the Markov states at the times at which the events were generated, i.e.

$$C_\ell = S(t_\ell)\,,$$

for $\ell = 1, \cdots, r$. Then the sequence $\{(C_\ell, Y_\ell), \ell = 1, \cdots, r\}$ is a *Markov renewal sequence* with transition density matrix

$$\exp\{(Q-\Lambda)y\}\Lambda\,,$$

where the jkth element is

$$\Pr\{C_\ell = k, Y_\ell = y \,|\, C_{\ell-1} = j\}$$

for all ℓ. Given conditional independence, the likelihood function is

$$\begin{aligned}\Pr\{Y^{(r)} = y^{(r)}\} &= \delta^{(0)}\left(\prod_{\ell=1}^{r}\exp\{(Q-\Lambda)y_\ell\}\Lambda\right)1' \\ &= \delta^{(0)}\exp\{(Q-\Lambda)y_1\}\Lambda\cdots\exp\{(Q-\Lambda)y_r\}\Lambda 1'\,, \end{aligned} \tag{4}$$

where $\delta^{(0)}$ are the state probabilities at $t = 0$.

2.4.2 Forward and Backward Equations

The forward equations (analogue of the discrete case in Eq 1) are

$$(\alpha_{\ell 1}, \cdots, \alpha_{\ell m}) = \begin{cases}\delta^{(0)} & \ell = 0 \\ \delta^{(0)}\exp\{(Q-\Lambda)y_1\}\Lambda\cdots\exp\{(Q-\Lambda)y_\ell\}\Lambda & \ell = 1, \cdots, r\,;\end{cases} \tag{5}$$

and the backward equations (analogue to those in Eq 2) are

$$(\beta_{\ell 1}, \cdots, \beta_{\ell m})' = \begin{cases} \exp\{(Q-\Lambda)y_{\ell+1}\}\Lambda \cdots \exp\{(Q-\Lambda)y_r\}\Lambda 1' & \ell = 0, \cdots, r-1 \\ 1' & \ell = r\,. \end{cases} \tag{6}$$

Hence, as in Eq 3,

$$L = (\alpha_{\ell 1}, \cdots, \alpha_{\ell m})(\beta_{\ell 1}, \cdots, \beta_{\ell m})', \quad \ell = 0, \cdots, r\,. \tag{7}$$

The matrices `alpha` and `beta` in the R function `HiddenMarkov::Estep.mmpp` contain

$$\begin{pmatrix} \alpha_{01} & \alpha_{02} & \cdots & \alpha_{0m} \\ \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{rm} \end{pmatrix} \text{ and } \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{pmatrix}, \tag{8}$$

respectively. In particular, note that the top row of `alpha` is $\delta^{(0)}$, and the bottom row of `beta` is a vector of ones. These matrices originate from `HiddenMarkov::forwardback.mmpp`.

2.5 EM Algorithm

2.5.1 Complete Data Likelihood

State transitions occur at τ_i, $i = 1, \cdots, n$, i.e.

$$S(\tau_i^-) \neq S(\tau_i)\,.$$

The sequence of visited Markov states is $\{S_i\}$ where $S_i = S(\tau_i)$. Then for $j \neq k$

$$\Pr\{S_i = k \,|\, S_{i-1} = j\} = \frac{q_{jk}}{q_j}\,.$$

The time in state S_i is $X_i = \tau_{i+1} - \tau_i$.

Note that we observe the process on $[0, \tau_{n+1})$, but not at τ_{n+1}. It follows that

$$\begin{aligned} &\Pr\{X^{(n)} = x^{(n)}, S^{(n)} = s^{(n)}\} \\ &= \Pr\{S_1 = s_1\} \left(\prod_{i=1}^{n-1} f_{X_i}(x_i \,|\, S_i = s_i) \Pr\{S_{i+1} = s_{i+1} \,|\, S_i = s_i\} \right) \times \\ &\qquad \Pr\{X_n \geq x_n \,|\, S_n = s_n\} \\ &= \delta_{s_1}^{(0)} \frac{q_{s_1 s_2}}{q_{s_1}} \frac{q_{s_2 s_2}}{q_{s_2}} \cdots \frac{q_{s_{n-1} s_n}}{q_{s_{n-1}}} \left(\prod_{i=1}^{n-1} q_{s_i} \exp(-q_{s_i} x_i) \right) \exp(-q_{s_n} x_n)\,. \end{aligned}$$

Note that this expression has the same form as in the discrete time case except for the last term, i.e. $\Pr\{X_n \geq x_n \,|\, S_n = s_n\}$. Further simplification gives

$$\Pr\{X^{(n)} = x^{(n)}, S^{(n)} = s^{(n)}\} = \delta^{(0)}_{s_1} q_{s_1 s_2} q_{s_2 s_3} \cdots q_{s_{n-1} s_n} \prod_{i=1}^{n} \exp(-q_{s_i} x_i)\,.$$

As in the discrete time case, define u_{ij} and v_{ijk} as:

$$u_{ij} = \begin{cases} 1 & \text{if } s_i = j \\ 0 & \text{otherwise} \end{cases}$$

$$v_{ijk} = \begin{cases} 1 & \text{if } s_{i-1} = j \text{ and } s_i = k \\ 0 & \text{otherwise}\,. \end{cases}$$

Then

$$\begin{aligned} &\log \Pr\{X^{(n)} = x^{(n)}, S^{(n)} = s^{(n)}\} \\ &\quad = \sum_{j=1}^{m} u_{1j} \log \delta^{(0)}_j + \sum_{j=1}^{m} \sum_{\substack{k=1 \\ k \neq j}}^{m} \left(\sum_{i=2}^{n} v_{ijk} \right) \log q_{jk} - \sum_{j=1}^{m} \sum_{i=1}^{n} u_{ij} x_i q_j\,. \end{aligned}$$

2.5.2 Add Event Times

The observed Poisson event times are denoted as $t_1, t_2, \cdots, t_r$, where r is the total number of events. We will denote the collection of all event times as $t^{(r)} = (t_1, t_2, \cdots, t_r)$.

Given we know $x^{(n)}$ and $s^{(n)}$, we can deduce the interval to which each event belongs, i.e. events in interval i are those with times in $\{t_\ell : \tau_i \leq t_\ell < \tau_{i+1}\}$. A useful variable for notational purposes is the number of events in the ith interval, which we will denote as r_i. Hence the event times in the ith interval can now be more explicitly stated as:

$$t_{\ell_i+1}, t_{\ell_i+2}, \cdots, t_{\ell_i+r_i},$$

where $\ell_i = \sum_{j=1}^{i-1} r_j$. However the "complete data" description of the process can be given as simply $(c^{(n)}, x^{(n)}, t^{(r)})$.

We want the joint density of inter-event times in the ith interval. Let the number of events in previous intervals be $\ell_i = \sum_{j=1}^{i-1} r_j$, then we want the joint density of these inter-event times:

$$t_{\ell_i+1} - \tau_i,\ t_{\ell_i+2} - t_{\ell_i+1},\ t_{\ell_i+3} - t_{\ell_i+2},\ \cdots,\ t_{\ell_i+r_i} - t_{\ell_i+r_i-1}$$

together with the density that no events occur in the interval $(t_{\ell_i+r_i}, \tau_{i+1})$. Denote this as

$$\begin{aligned} &f(\tau_i, t_{\ell_i+1}, t_{\ell_i+2}, \cdots, t_{\ell_i+r_i}, \tau_{i+1} \,|\, S_i = s_i, X_i = x_i) \\ &\quad = \lambda_{s_i} \exp[-\lambda_{s_i}(t_{\ell_i+1} - \tau_i)]\ \lambda_{s_i} \exp[-\lambda_{s_i}(t_{\ell_i+2} - t_{\ell_i+1})]\ \cdots \\ &\qquad\qquad \lambda_{s_i} \exp[-\lambda_{s_i}(t_{\ell_i+r_i} - t_{\ell_i+r_i-1})] \exp[-\lambda_{s_i}(\tau_{i+1} - t_{\ell_i+r_i})] \\ &\quad = \lambda_{s_i}^{r_i} \exp(-\lambda_{s_i}(\tau_{i+1} - \tau_i)) \\ &\quad = \lambda_{s_i}^{r_i} \exp(-\lambda_{s_i} x_i) \\ &\quad = \frac{(\lambda_{s_i} x_i)^{r_i}}{r_i!} \exp(-\lambda_{s_i} x_i) \frac{r_i!}{x_i^{r_i}}\,, \end{aligned}$$

where λ_{s_i} is the Poisson event rate while the Markov process is in state s_i. Taking logarithms and summing over all visited Markov states, we get

$$\begin{aligned}
\sum_{i=1}^{n} \log f(\tau_i, t_{\ell_i+1}, t_{\ell_i+2}, \cdots, t_{\ell_i+r_i}, \tau_{i+1} \,|\, S_i = s_i, X_i = x_i) \\
&= \sum_{i=1}^{n} r_i \log \lambda_{s_i} - \sum_{i=1}^{n} x_i \lambda_{s_i} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij} r_i \log \lambda_j - \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij} x_i \lambda_j \,.
\end{aligned}$$

Note that the last subinterval ($i = n$) terminates at t_r.

The "complete data" log-likelihood is then

$$\begin{aligned}
\log \Pr\{X^{(n)} = x^{(n)}, S^{(n)} = s^{(n)}, T^{(r)} = t^{(r)}\} \\
&= \sum_{j=1}^{m} u_{1j} \log \delta_j^{(0)} + \sum_{j=1}^{m} \sum_{\substack{k=1 \\ k \neq j}}^{m} \left(\sum_{i=2}^{n} v_{ijk} \right) \log q_{jk} - \sum_{j=1}^{m} \sum_{i=1}^{n} u_{ij} x_i (q_j + \lambda_j) + \\
&\qquad \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij} r_i \log \lambda_j \,.
\end{aligned}$$

Rydén (1996, Eq 7) uses a different notation. However, the terms in common are

$$\begin{aligned}
\text{time spent in state } j &= \sum_{i=1}^{n} u_{ij} x_i \,, \\
\text{number of events occurring in state } j &= \sum_{i=1}^{n} u_{ij} r_i \,, \text{ and} \\
\text{number of switches from state } j \text{ to } k &= \sum_{i=2}^{n} v_{ijk} \,.
\end{aligned}$$

2.5.3 E-Step and M-Step

The model parameters that require estimation are $\Theta = (Q, \Lambda)$. To implement the EM algorithm, one initially needs to analytically evaluate the *expectations*:

$$\mathrm{E}\left[\sum_{i=1}^{N} U_{ij} X_i \,\middle|\, T^{(r)} = t^{(r)} \right], \quad \mathrm{E}\left[\sum_{i=1}^{N} U_{ij} R_i \,\middle|\, T^{(r)} = t^{(r)} \right], \text{ and } \mathrm{E}\left[\sum_{i=2}^{N} V_{ijk} \,\middle|\, T^{(r)} = t^{(r)} \right],$$

where the uppercase variables within the expectations are the corresponding random variables to the lower case realisations. One then uses these expressions as estimators; and together with the current parameter estimates $\widehat{\Theta}$, the terms (i.e. "missing data")

$$\sum_{i=1}^{n} u_{ij} x_i \,, \quad \sum_{i=1}^{n} u_{ij} r_i \,, \text{ and } \quad \sum_{i=2}^{n} v_{ijk}$$

are estimated. This is referred to as the *expectation* or *E-step*.

These values then replace the corresponding terms in the complete data likelihood. Then new values are estimated for $\widehat{\Theta}$ by *maximising* this complete data likelihood. This is referred to as the *maximisation* or *M-step*. The process is repeated until the estimates $\widehat{\Theta}$ converge.

Evaluation of the above expectations pose a number of problems, both analytical and numerical. This appears to be the most complicated aspect in the application of the EM algorithm to the MMPP model. Rydén (1996) derives expressions for the expectations based on an eigenvalue decomposition, while the expressions derived by Klemm et al (2003) use a Poisson like series expansion.

2.5.4 Addition of Marks

The complication mentioned above is further compounded by the addition of "marks". Let $W^{(r)}$ be marks associated with each of the r events. Assume that the marks have an exponential distribution with rate parameter ξ_j when the Markov process $S(t)$ is within state j. Let

$$\Xi = \text{diag}[\xi_1, \cdots, \xi_m] \,,$$

so now the model parameters that require estimation are $\Theta = (Q, \Lambda, \Xi)$.

The joint density of the marks in the ith interval is

$$f(w_{\ell_i+1}, w_{\ell_i+2}, \cdots, w_{\ell_i+r_i} \,|\, S_i = s_i, X_i = x_i) = \xi_{s_i}^{r_i} \exp\left(-\xi_{s_i} \sum_{\ell=\ell_i+1}^{r_i} w_\ell \right) \,,$$

where $\ell_i = r_1 + \cdots + r_{i-1}$. This will add two further terms to the complete data likelihood:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij} r_i \log \xi_j - \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij} \xi_j \sum_{\ell=\ell_i+1}^{r_i} w_\ell \,.$$

Hence, the "complete data" log-likelihood is now

$$\begin{aligned} & \log \Pr\{X^{(n)} = x^{(n)}, S^{(n)} = s^{(n)}, T^{(r)} = t^{(r)}, W^{(r)} = w^{(r)}\} \\ &= \sum_{j=1}^{m} u_{1j} \log \delta_j^{(0)} + \sum_{j=1}^{m} \sum_{\substack{k=1 \\ k \neq j}}^{m} \left(\sum_{i=2}^{n} v_{ijk} \right) \log q_{jk} - \sum_{j=1}^{m} \sum_{i=1}^{n} u_{ij} x_i (q_j + \lambda_j) \\ & \qquad + \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij} r_i \log(\xi_j \lambda_j) - \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{\ell=\ell_i+1}^{r_i} u_{ij} w_\ell \xi_j \,. \end{aligned}$$

How do these new terms affect the expectations in §2.5.3. Now our observed data are $T^{(r)}$ and $W^{(r)}$, and so the expectations are now conditional on both, i.e. we want

$$\text{E}\left[\sum_{i=1}^{N} U_{ij} X_i \,\middle|\, T^{(R)} = t^{(r)}, W^{(R)} = w^{(r)} \right] \,, \quad \text{E}\left[\sum_{i=1}^{N} U_{ij} R_i \,\middle|\, T^{(R)} = t^{(r)}, W^{(R)} = w^{(r)} \right] \,,$$

and

$$\mathrm{E}\left[\sum_{i=2}^{N} V_{ijk} \,\middle|\, T^{(R)} = t^{(r)}, W^{(R)} = w^{(r)} \right] ,$$

together with the new term

$$\mathrm{E}\left[\sum_{i=1}^{N} \sum_{\ell=\ell_i+1}^{R_i} U_{ij} W_\ell \,\middle|\, T^{(R)} = t^{(r)}, W^{(R)} = w^{(r)} \right] ,$$

where $\ell_i = R_1 + \cdots + R_{i-1}$. The new term is the expected sum of marks "emitted" by events that occur while the Markov process is in state j.

2.6 Particle Filters

See Arulampalam et al (2002), Doucet et al (2001), and Doucet & Tadić (2003).

3 References

Arulampalam, M.S.; Maskell, S.; Gordon, N.J. & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* **50**(2), 174–188. DOI: 10.1109/78.978374

Doucet, A. & Tadić, V.B. (2003). Parameter estimation in general state-space models using particle methods. *Annals of the Institute of Statistical Mathematics* **55**(2), 409–422. DOI: 10.1007/BF02530508

Doucet, A.; Gordon, N.J. & Krishnamurthy, V. (2001). Particle filters for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing* **49**(3), 613–624. DOI: 10.1109/78.905890

Elliott, R.J.; Aggoun, L. & Moore, J.B. (1994). *Hidden Markov Models: Estimation and Control.* Springer-Verlag, New York. ISBN: 0-387-94364-1

Harte, D.S. (2010). *Package "HiddenMarkov": Hidden Markov Models.* R package version 1.4-0. Comprehensive R Archive Network (CRAN).

Klemm, A.; Lindemann, C. & Lohmann, M. (2003). Modeling IP traffic using the batch Markovian arrival process. *Performance Evaluation* **54(2)**, 149–173. DOI: 10.1016/S0166-5316(03)00067-1

MacDonald, I.L. & Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman and Hall/CRC, Boca Raton. ISBN: 0-412-55850-5

Meier-Hellstern, K.S. (1987). A fitting algorithm for Markov-modulated poisson processes having two arrival rates. *European Journal of Operations Research* **29(3)**, 370–377. DOI: 10.1016/0377-2217(87)90250-5

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, ISBN 3-900051-07-0. URL: www.r-project.org/.

Rao, A.R. & Hamed, K.H. (2000). *Flood Frequency Analysis.* CRC, Boca Raton. ISBN: 0-412-55280-9

Roberts, W.J.J.; Ephraim, Y. & Dieguez, E. (2006). On Rydén's EM algorithm for estimating MMPPs. *IEEE Signal Processing Letters* **13(6)**, 373–376. DOI: 10.1109/LSP.2006.871709

Rydén, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis* **21(4)**, 431–447. DOI: 10.1016/0167-9473(95)00025-9

Zucchini, W. (2005). *Hidden Markov Models Short Course, 3–4 April 2005.* Macquarie University, Sydney.