An EM algorithm for Markovian arrival processes observed at discrete times

Lothar Breuer, Alfred Kume
University of Kent, Canterbury, UK

Abstract. The present paper contains a specification of the EM algorithm in order to fit an empirical counting process, observed at discrete times, to a Markovian arrival process. The given data are the numbers of observed events in disjoint time intervals. The underlying phase process is not observable. An exact numerical procedure to compute the E and M steps is given.

1. Introduction

Markovian arrival processes have been introduced by Neuts (1979) and Lucantoni (1991). They have extensively been used as models for input streams to queueing systems (for a survey see Lucantoni (1993)). Their appealing feature is that they are Markovian (and hence analytically tractable) on the one hand but very versatile (even dense in the class of point processes, see Asmussen and Koole (1993)) on the other hand. Although the concept of Markovian arrival processes (MAPs†) has gained widespread use in stochastic modelling of communication systems and other application areas, the quest for the best statistical methods of parameter estimation is far from finished yet.

A survey of estimation methods is given in Asmussen (1997). His emphasis is on maximum likelihood estimation and its implementation via the EM algorithm (see Dempster et al. (1977)). Asmussen et al. (1996) derived a fitting procedure for phase-type distributions via the EM algorithm. Markov chain Monte Carlo methods for the estimation of phase-type distributions (and functionals of these) are given in Bladt et al. (2003). For a special case of MAPs, the Markov–modulated Poisson Process (MMPP), an EM algorithm has been developed in Ryden (1996). Bladt and Soerensen (2005) specify the EM algorithm for the case of discretely observed Markov jump processes (MJPs). We will have to deal with discretely observed MJPs, for which even the observations at discrete times are partial only. Fearnhead and Sherlock (2006) provide a simulation method for MMPPs. Our results extend this paper in so far in that we provide a maximum likelihood approach for a more general class of processes.

Statistical model fitting depends of course on the type of data observation that is available. In practice, we think of essentially two types of data:
(a) Exact times are recorded for each observed event.
(b) The arrival process is observed at a grid of discrete times only. This yields only the information of how many arrivals have occurred in each interval of the grid.
We always assume that the underlying phase process is unobservable. An EM algorithm for case (a) has been given in Breuer (2002). The present paper contains a specification of the EM algorithm for MAPs in the case (b) of discrete time observation.

†There is some confusion in the literature about this acronym. It is used also for Markov–additive processes, which form a much more general class than Markovian arrival processes. However, since MAP is the most common abbreviation for Markovian arrival processes, we will use it in this article.

In section 2, we review shortly the main definitions and notations for MAPs. The EM algorithm is specified to discretely observed MAPs with hidden phases in section 3. Exact expressions for the integrals appearing in the E–step are given in sections 4 and 5. In the remainder of this section, we describe the kind of data available from observations and give a short remark on estimating the order of the MAP which is to fit the empirical time series.

We assume to be in the following, for many applications typical, situation of data retrieval: An empirical counting process is observed at discrete times t_n, $n = 0, \ldots, N$, where $t_0 := 0$. To simplify later notation we assume that $t_n := n$. It will be apparent that this assumption of equidistant observation points is not necessary. We further assume that the observed point process is stationary in time. The only information that can be measured is the number of observed events in the interval $]t_{n-1}, t_n]$, denoted by z_n, $n = 1, \ldots, N$. Thus the given data has the form $z = (z_1, \ldots, z_N)$. Due to the result that MAPs are dense in the class of all point processes on the positive real axis (see Asmussen and Koole (1993)), the approach of model fitting by a MAP is reasonable. By the nature of the problem, no information is given on the underlying phase process, not even the number of phases.

Throughout this article, we fix the number of phases for the MAP model to be a known integer $m \geq 1$. Procedures for estimating the number m of phases are discussed in Ryden (1997) for the MMPP.

Since the adaptation of the model increases with the assumed number of phases m, the likelihood gain at the ML estimates is always positive if we increase m by 1. If this gain is not bigger than some threshold value, we can assume that we have found the right value for m. This incremental method was proposed by Jewell (1982). The threshold value reflects the limit of accuracy beyond which the gain in model adaptation is not worth the additional computation time.

2. Markovian arrival processes

A Markovian arrival process is a homogeneous Markov process $\mathcal{Y} = (Y_t : t \geq 0)$ with state space $E = \mathbb{N}_0 \times \{1, \ldots, m\}$, where m is some positive integer, and a generator matrix of the (block) form

$$G = \begin{pmatrix} D_0 & D_1 & & \\ & D_0 & D_1 & \\ & & \ddots & \ddots \end{pmatrix}$$

In this generator, only the main and the first upper block diagonals have non–zero entries. Apart from that there are no restrictions for the matrices D_0 and D_1, except of course the generator conditions

$$(D_0 + D_1)\mathbf{1} = \mathbf{0}, \quad D_{0;ij} \geq 0 \quad \text{for } i \neq j, \quad D_{1;ij} \geq 0 \quad \text{for all } i, j$$

where $\mathbf{1}$ and $\mathbf{0}$ denote the vectors with all entries being ones and zeroes, respectively. In order to avoid absorbing states, we assume that $D_{0;ii}$ is strictly negative for all $i = 1, \ldots, m$.

As Y_t is two–dimensional, it is natural to write $\mathcal{Y} = (\mathcal{N}, \mathcal{J}) = ((N_t, J_t) : t \geq 0)$. The marginal processes $\mathcal{N}$ and $\mathcal{J}$ are called the counting process and the phase process associated with $\mathcal{Y}$. For any state $(n, i) \in E$, the first dimension is called the level and the second one is called the phase.

Under this interpretation, the entries $D_{0;ij}$ of D_0 give the infinitesimal transition rates among phases $\{1, \ldots, m\}$ without an arrival if $i \neq j$. The entries $D_{1;ij}$ of D_1 give the infinitesimal transition rates among phases accompanied by an arrival. The diagonal entries $D_{0;ii}$ are the negative parameters of the exponential sojourn times in any state (n, i), independently from $n \in \mathbb{N}_0$.

The special case of a Markov–modulated Poisson Process (MMMP), also called a Cox process, arises if D_1 is chosen to be a diagonal matrix. This has of course a large impact on the modelling power. The matrix D_1 governs correlations between inter–arrival times (which are crucial in many applications). If D_1 is restricted to be diagonal, there is no way to control these. This is the main reason why MMPPs can be employed only in special modelling situations.

The process has stationary increments if it starts in phase equilibrium π, which is determined as the stationary distribution of the phase process, i.e. by $\pi(D_0 + D_1) = \mathbf{0}$. As we wish to fit a stationary empirical point process by a MAP, we can hence assume that $\mathbb{P}(Y_0 = (0, i)) = \pi_i$.

The likelihood of a complete sample path x on $[0, t_N]$ under parameters $D_0 = (D_{0;ij})$ and $D_1 = (D_{1;ij})$ is given by

$$f(x|D_0, D_1) = \prod_{i=1}^{m} \exp(D_{0;ii} Z_i) \prod_{i=1}^{m} \prod_{j=1, j\neq i}^{m} D_{0;ij}^{B_{ij}} \prod_{i=1}^{m} \prod_{j=1}^{m} D_{1;ij}^{A_{ij}} \tag{1}$$

where Z_i denotes the total time spent in phase i, B_{ij} the number of jumps from phase i to phase j without arrival, and A_{ij} the number of jumps from phase i to phase j with accompanying arrival. These variables form a sufficient statistic for likelihood based estimations. They can of course be decomposed into the sum of the respective variables over all the intervals $]t_{n-1}, t_n]$, $n = 1, \ldots, N$. Thus we can write

$$Z_i = \sum_{n=1}^{N} Z_i^n, \qquad B_{ij} = \sum_{n=1}^{N} B_{ij}^n, \qquad A_{ij} = \sum_{n=1}^{N} A_{ij}^n$$

where Z_i^n, B_{ij}^n, and A_{ij}^n refer to the nth interval.

Remark:
The above equation (1) shows that under the complete statistics

$$T(x) = (Z_i : i \leq m; B_{ij} : i \neq j; A_{ij} : i, j \leq m)$$

we are dealing with an exponential family in T and a parameter vector

$$\zeta(D_0, D_1) = (D_{0;ii} : i \leq m; \log D_{0;ij} : i \neq j; \log D_{1;ij} : i, j \leq m)^T$$

where we define by natural extension $log\ 0 := -\infty$ and $(-\infty) \cdot 0 := 0$. Under this setting we obtain

$$f(x|D_0, D_1) = \exp(T(x)\zeta(D_0, D_1))$$

This shows that results obtained by Sundberg (1974) are applicable to the problem studied here.

Acknowledging the relation $D_{0;ii} = -\left(\sum_{j=1}^{m} D_{1;ij} + \sum_{j=1, j\neq i}^{m} D_{0;ij}\right)$, the maximum likelihood estimators $\hat{D}_0$ and $\hat{D}_1$ for the matrices D_0 and D_1 are then given by

$$\hat{D}_{0;ij} = \frac{B_{ij}}{Z_i}, \qquad \hat{D}_{1;ij} = \frac{A_{ij}}{Z_i}, \tag{2}$$

$$\hat{D}_{0;ii} = -\left(\sum_{j=1}^{m} \hat{D}_{1;ij} + \sum_{j=1, j\neq i}^{m} \hat{D}_{0;ij}\right) \tag{3}$$

for $1 \leq i, j \leq m$, see Albert (1962).

3. The EM algorithm

The typical property of observing time series derived from a MAP is that only the arrivals but not the phases can be seen. If the phases were observable, then one could apply the maximum likelihood estimators for finite state Markov processes (see Albert (1962)). To make things worse, we cannot even observe the exact arrival times. Thus we have a problem of estimation from incomplete data. For this type of statistical problems, the so–called EM algorithm has proven to be a good means of approximating the maximum likelihood estimator (see Dempster et al. (1977), McLachlan and Krishnan (1997) or Meng and van Dyk (1997)). The name EM algorithm stems from the alternating application of an expectation step (for E) and a maximization step (for M) which yield successively higher likelihoods of the estimated parameters.

In our case, the incomplete sample consists only of the sequence $z = (z_1, \ldots, z_N)$ indicating the number of observed arrivals within each interval. Denote the the maximal observed number of arrivals within one interval by M.

Given the parameters D_0 and D_1 as well as the stationary phase distribution π (which is determined by $D_0 + D_1$), the likelihood of the incomplete sample z is

$$f(z|D_0, D_1) = \pi \prod_{n=1}^{N} g(z_n|D_0, D_1)\mathbf{1} \tag{4}$$

with $g(0|D_0, D_1) = e^{D_0}$ and

$$g(i|D_0, D_1) = \int\limits_{u_0+\ldots+u_i=1} \left(\prod_{n=0}^{i-1} e^{D_0 u_n} D_1 \right) e^{D_0 u_i} \, du_0 \ldots du_{i-1}$$

for $i \geq 1$.

Assume that the estimates after the kth EM iteration are given by the matrices $(\hat{D}_0^{(k)}, \hat{D}_1^{(k)})$. Then in the first step of the $k+1$st cycle, the conditional expectations of the variables Z_i, A_{ij} and B_{ij} given the incomplete observation and the current estimates $(\hat{D}_0^{(k)}, \hat{D}_1^{(k)})$ are computed.

In order to simplify notations, define the column vectors

$$\eta_N := \mathbf{1} \qquad \text{and} \qquad \eta_{n-1} := g(z_n|\hat{D}_0^{(k)}, \hat{D}_1^{(k)}) \, \eta_n = \prod_{i=n}^{N} g(z_i|\hat{D}_0^{(k)}, \hat{D}_1^{(k)})\mathbf{1} \tag{5}$$

iteratively for $2 \leq n \leq N$.

Since the empirical time series is observed in a stationary regime, we can set the phase distribution α_0 at time 0 to be the estimated phase equilibrium, i.e. satisfying $\alpha_0(\hat{D}_0^{(k)} + \hat{D}_1^{(k)}) = 0$. Then we define iteratively the row vectors

$$\alpha_{n+1} := \alpha_n \, g(z_{n+1}|\hat{D}_0^{(k)}, \hat{D}_1^{(k)}) = \pi \prod_{i=0}^{n} g(z_i|\hat{D}_0^{(k)}, \hat{D}_1^{(k)}) \tag{6}$$

for $0 \leq n \leq N-2$. Clearly $f(z|\hat{D}_0^{(k)}, \hat{D}_1^{(k)} = \alpha_n \eta_n$.

We begin the E–step with the accumulated sojourn times in a phase i. These are given by

$$Z_i^{(k+1)} := \mathbb{E}_{(\hat{D}_0^{(k)}, \hat{D}_1^{(k)})}(Z_i|z) = \left(f(z|\hat{D}_0^{(k)}, \hat{D}_1^{(k)}) \right)^{-1} \sum_{n=1}^{N} E_{(\hat{D}_0^{(k)}, \hat{D}_1^{(k)})}(Z_i^n|z)$$

where $i = 1 \leq m$ and Z_i^n denotes the random variable of the total amount of time spent in phase i within the nth interval. The terms in the sum are given by

$$E_{(\hat{D}_0^{(k)},\hat{D}_1^{(k)})}(Z_i^n|z) = \alpha_{n-1}\, c_{z_n}(i,i|\hat{D}_0^{(k)},\hat{D}_1^{(k)})\, \eta_n \tag{7}$$

for all $0 \leq n \leq N$, where the matrix–valued functions c_n are defined as

$$c_0(i,j|\hat{D}_0^{(k)},\hat{D}_1^{(k)}) := \int_0^1 \exp(\hat{D}_0^{(k)}u)e_i \cdot e_j^T \exp(\hat{D}_0^{(k)}(1-u))\, du \tag{8}$$

$$\begin{aligned} c_n(i,j|\hat{D}_0^{(k)},\hat{D}_1^{(k)}) := \int\limits_{u_0+\ldots+u_n=1} \sum_{h=0}^{n} \left(\prod_{l=0}^{h-1} \exp(\hat{D}_0^{(k)}u_l)\hat{D}_1^{(k)}\right) \\ \int_0^{u_h} \exp(\hat{D}_0^{(k)}v)e_i \cdot e_j^T \exp(\hat{D}_0^{(k)}(u_h-v))\, dv \\ \left(\prod_{l=h+1}^{n} \hat{D}_1^{(k)} \exp(\hat{D}_0^{(k)}u_l)\right) du_0 \ldots du_{n-1} \end{aligned} \tag{9}$$

for $1 \leq i,j \leq m$ and $1 \leq n \leq M$. Here e_i denotes the ith canonical column base vector and e_i^T its transpose, i.e. the row vector. The empty products $\prod_{l=0}^{-1} \ldots$ and $\prod_{l=n+1}^{n} \ldots$ are defined as the identity matrix. The values for $c_n(i,j|\hat{D}_0^{(k)},\hat{D}_1^{(k)})$ can be rewritten in terms of the $n+2$-dimensional simplex as

$$\begin{aligned} \sum_{h=0}^{n} \int\limits_{u_0+\ldots+u_{n+1}=1} \left(\prod_{l=0}^{h-1} \exp(\hat{D}_0^{(k)}u_l)\hat{D}_1^{(k)}\right) \exp(\hat{D}_0^{(k)}u_h)e_i \cdot e_j^T \exp(\hat{D}_0^{(k)}u_{h+1}) \\ \left(\prod_{l=h+2}^{n+1} \hat{D}_1^{(k)} \exp(\hat{D}_0^{(k)}u_l)\right) du_0 \ldots du_h \ldots du_n \end{aligned} \tag{10}$$

The derivation of (7) is completely analogous to the one in Asmussen et al. (1996), p.439. Likewise,

$$B_{ij}^{(k+1)} := E_{(\hat{D}_0^{(k)},\hat{D}_1^{(k)})}(B_{ij}|z) = \left(f(z|\hat{D}_0^{(k)},\hat{D}_1^{(k)})\right)^{-1} \sum_{n=1}^{N} E_{(\hat{D}_0^{(k)},\hat{D}_1^{(k)})}(B_{ij}^n|z)$$

with

$$E_{(\hat{D}_0^{(k)},\hat{D}_1^{(k)})}(B_{ij}^n|z) = \hat{D}_{0;ij}^{(k)} \cdot \alpha_{n-1}\, c_{z_n}(i,j|\hat{D}_0^{(k)},\hat{D}_1^{(k)})\, \eta_n \tag{11}$$

for $1 \leq n \leq N$ is derived using completely the same arguments as in Asmussen et al. (1996), p.440. The E–step is completed by

$$A_{ij}^{(k+1)} := E_{(\hat{D}_0^{(k)},\hat{D}_1^{(k)})}(A_{ij}|z) = \left(f(z|\hat{D}_0^{(k)},\hat{D}_1^{(k)})\right)^{-1} \sum_{n=1}^{N} E_{(\hat{D}_0^{(k)},\hat{D}_1^{(k)})}(A_{ij}^n|z)$$

with

$$E_{(\hat{D}_0^{(k)},\hat{D}_1^{(k)})}(A_{ij}^n|z) = \begin{cases} 0, & z_n = 0 \\ \hat{D}_{1;ij}^{(k)} \cdot \alpha_{n-1}\, c_{z_n-1}(i,j|\hat{D}_0^{(k)},\hat{D}_1^{(k)})\, \eta_n, & z_n > 0 \end{cases} \tag{12}$$

for $1 \leq n \leq N$.

Remark

Based on the fact that

$$\frac{\partial \exp(D_0^{(k)} u_h)}{\partial D_{0;ij}^{(k)}} = u_h \int_0^1 \exp(tD_0^{(k)} u_h) e_i e_j^T \exp((1-t)D_0^{(k)} u_h) dt$$
$$= \int_0^{u_h} \exp(D_0^{(k)} v) e_i e_j^T \exp(D_0^{(k)}(u_h - v)) dv$$

and by transferring the derivative inside the integration sign one can easily see that

$$\frac{\partial g(n|D_0^{(k)}, D_1^{(k)})}{\partial D_{0;ij}^{(k)}} = c_n(i,j|D_0^{(k)}, D_1^{(k)}) \quad \text{and} \quad \frac{\partial g(n|D_0^{(k)}, D_1^{(k)})}{\partial D_{1;ij}^{(k)}} = c_{n-1}(i,j|D_0^{(k)}, D_1^{(k)}).$$

Therefore

$$\frac{\partial f(z|D_0^{(k)}, D_1^{(k)})}{\partial D_{0;ij}^{(k)}} = \sum_{r=1}^{N} \alpha_{r-1} c_{z_r}(i,j|, D_0^{(k)}, D_1^{(k)}) \eta_r$$

and

$$\frac{\partial f(z|D_0^{(k)}, D_1^{(k)})}{\partial D_{1;ij}^{(k)}} = \sum_{r=1}^{N} \alpha_{r-1} c_{z_r-1}(i,j|D_0^{(k)}, D_1^{(k)}) \eta_r.$$

Similarly, one could easily obtain the second or higher order derivatives of the likelihood. These derivatives can be easily adopted for the cases when there is some simple functional relationship between the entries of the parameter matrices $D_0^{(k)}$ and $D_1^{(k)}$.

Now, the next step of the $k+1$st cycle of the EM consists of the computation of maximum likelihood estimates given the new (conditional but complete) statistic computed in the E–step. This can be done by simply replacing the variables in equations (2) and (3) by the conditional expectations computed above. This leads to re-evaluated estimates

$$\hat{D}_{0;ij}^{(k+1)} = \frac{B_{ij}^{(k+1)}}{Z_i^{(k+1)}}, \qquad \hat{D}_{1;ij}^{(k+1)} = \frac{A_{ij}^{(k+1)}}{Z_i^{(k+1)}},$$

and

$$\hat{D}_{0;ii}^{(k+1)} = -\left(\sum_{j=1}^{m} \hat{D}_{1;ij}^{(k+1)} + \sum_{j=1, j\neq i}^{m} \hat{D}_{0;ij}^{(k+1)} \right)$$

for $1 \leq i, j \leq m$.

Using these, one can compute the likelihood $f(z|\hat{D}_0^{(k+1)}, \hat{D}_1^{(k+1)})$ of the empirical time series under the new estimates according to equation (4). If the likelihood ratio

$$\rho = \frac{f(z|\hat{D}_0^{(k+1)}, \hat{D}_1^{(k+1)})}{f(z|\hat{D}_0^{(k)}, \hat{D}_1^{(k)})}$$

remains smaller than a threshold $1 + \varepsilon$, then the EM iteration process can be stopped, and the latest estimates may be adopted. The threshold value reflects the limit of accuracy beyond which the gain in model adaptation is considered not to be worth the additional computation time.

4. Implementation of EM

In this section we focus on the implementation of the EM algorithm.

It is clear that in order to evaluate the update rules we need to be able to calculate the value of the matrix integral of the type

$$\mathcal{I} = \int\limits_{u_0+\cdots+u_k=1} e^{D_0 u_0} P_1 e^{D_0 u_1} \cdots P_k e^{D_0 u_k} d\mathbf{u}$$

where D_0 is our parameter square matrix of order m and P_r's are equal to D_1 except for calculating the elementary terms for $c_{k-1}(i,j|\hat{D}_0,\hat{D}_1)$ in (10) where one P_r needs to replaced with $e_i e_j^T$. Note that the true likelihood function is also constructed in terms of such expectations.

In the following, we show two ways to calculate $\mathcal{I}$. The first operates with the matrices of the same dimension as D_0 and D_1 and the second is based on the matrix exponentials of some large dimension depending on k and m. The choice of these approaches for practical implementation will depend on the computing limitations related to large values of m and k.

4.1. Direct evaluation

The purpose of this sub-section is two fold:

First, to demonstrate a direct method of evaluating the density function of a convolution of Erlang distributions. To our knowledge this is not reported before and we show it to be closely related to the normalizing constant of a particular spherical distribution.

Secondly, the expression of the density function of the convolutions of Erlang's is shown to be closely related to the close form expression of $\mathcal{I}$. In particular, if the Jordan decomposition of D_0 is known, we have a closed form expression for evaluating both the likelihood function and the EM update steps.

4.1.1. Convolutions of Erlang distributions

Let assume that the random variables X_i have distribution $Gamma(n_i+1,\lambda_i)$ for $i=0,\cdots,k$. Since we will obtain the pdf of $Y=X_0+X_1+\cdots+X_k$, we can assume without loss of generality that all λ_i's are distinct.

It is now clear that the pdf of Y at point s is

$$f_Y(s) = \prod_{i=0}^{k} \frac{\lambda_i^{n_i+1}}{n_i!} \int\limits_{v_0+\cdots+v_k=s} e^{-\sum_{i=0}^{k} v_i\lambda_i} \prod_{i=0}^{k} v_i^{n_i} d\mathbf{v}$$

where $d\mathbf{v} = \prod_{i=1}^{k} dv_i$. A change of variables $u_i = v_i/s$ leads to

$$f_Y(s) = \prod_{i=0}^{k} \frac{\lambda_i^{n_i+1}}{n_i!} s^{k+\sum_{i=0}^{k} n_i} \int\limits_{u_0+\cdots+u_k=1} e^{-\sum_{i=0}^{k} u_i s\lambda_i} \prod_{i=0}^{k} u_i^{n_i} d\mathbf{u} \tag{13}$$

Integrals of the type $\int\limits_{u_0+\cdots+u_k=1} e^{-\sum_{i=0}^{k} u_i\lambda_i} \prod_{i=0}^{k} u_i^{n_i} d\mathbf{u}$ are obtained in the closed form in Kume and Wood (2007). The authors provide the value of the normalizing constant for Complex

Bingham Distributions (see Kent (1994)) with multiplicities in the eigenvalues of the parameter matrix. In particular, provided that all λ_i's are distinct, it is shown in Proposition 2 there that

$$\int_{u_0+\cdots+u_k=1} e^{-\sum_{i=0}^k u_i\lambda_i}\prod_{i=0}^k u_i^{n_i}d\mathbf{u} = \sum_{i=0}^k \sum_{|J_0(i)|=n_i} \frac{(-1)^{n_i+j}e^{-\lambda_i}n_i!}{j!j_0!\dots j_{i-1}!j_{i+1}!\dots j_k!}\prod_{r\neq i}\frac{(n_r+j_r)!}{(\lambda_r-\lambda_i)^{n_r+j_r+1}} \quad (14)$$

where the second summation is performed along all $J_0(i) = (j, j_0 \dots, j_{i-1}, j_{i+1}, \dots, j_k)$, which are integer partitions (including zeros) of n_i in $k+1$ components. A simple algorithm which generates such partitions is given on page 49 of Nijenhuis and Herbert (1978) where the number of such partitions is shown to be $C^{n_i}_{n_i+k}$. Replacing λ_i's in (14) by $s\lambda_i$'s we have

$$\int_{u_0+\cdots+u_k=1} e^{-\sum_{i=0}^k u_i s\lambda_i}\prod_{i=0}^k u_i^{n_i}d\mathbf{u} = s^{-k-\sum_{i=0}^k n_i}\sum_{i=0}^k \sum_{|J_0(i)|=n_i} \frac{(-1)^{n_i+j}e^{-s\lambda_i}s^j n_i!}{j!j_0!\dots j_{i-1}!j_{i+1}!\dots j_k!}\prod_{r\neq i}\frac{(n_r+j_r)!}{(\lambda_r-\lambda_i)^{n_r+j_r+1}} \quad (15)$$

which implies that

$$f_Y(s) = \prod_{i=0}^k \frac{\lambda_i^{n_i+1}}{n_i!}\sum_{i=0}^k e^{-s\lambda_i}\sum_{|J_0(i)|=n_i} \frac{(-1)^{n_i+j}s^j n_i!}{j!j_0!\dots j_{i-1}!j_{i+1}!\dots j_k!}\prod_{r\neq i}\frac{(n_r+j_r)!}{(\lambda_r-\lambda_i)^{n_r+j_r+1}} \quad (16)$$

The expression (14) is also valid for complex values of λ_i. This fact is important in calculating our expectations if the eigenvalues of D_0 are complex.

4.1.2. Evaluating $\mathcal{I}$

Assume that D_0 has p distinct eigenvalues with Jordan decomposition

$$D_0 = O\Delta O^{-1} = O\begin{pmatrix}\Delta_1(r_1) & & \\ & \ddots & \\ & & \Delta_p(r_p)\end{pmatrix}O^{-1} \quad \text{with} \quad \Delta_j(r_j) = \begin{pmatrix}\lambda_j & 1 & & \\ & \lambda_j & \ddots & \\ & & & 1 \\ & & & \lambda_j\end{pmatrix}$$

where r_j is the dimension of $\Delta_j(r_j)$ and O is an invertible matrix. Without loss of generality we can assume that O is the identity matrix. This can easily be seen from the fact that $e^{xD_0} = Oe^{x\Delta}O^{-1}$ and so

$$\mathcal{I} = O\int_{u_0+\cdots+u_k=1} e^{\Delta u_0}Q_1e^{\Delta u_1}\cdots Q_ke^{\Delta u_k}d\mathbf{u}\ O^{-1} \quad Q_j = O^{-1}D_jO.$$

We need to make the following remarks

Remark 1

$$e^{x\Delta} = \begin{pmatrix}e^{x\Delta_1(r_1)} & & \\ & \ddots & \\ & & e^{x\Delta_p(r_p)}\end{pmatrix}$$

Remark 2 From the decomposition $\Delta_j(r_j) = \lambda_j I + N(r_j)$ with

$$N(r_j) = \begin{pmatrix} & 1 & \\ & \ddots & \\ & & 1 \\ & & \end{pmatrix}$$

and noting that $N(r_j)$ is a nilpotent matrix of order r_j, it follows that

$$e^{x\Delta_j(r_j)} = e^{x\lambda_j} \sum_{w=0}^{r_j-1} \frac{x^w N(r_j)^w}{w!}.$$

with $N(r_j)^0 = I$.

Let denote by $M(j,w)$ the $p \times p$ matrix related to the block matrix $N(r_j)$ defined as

$$M(j,w) = \begin{pmatrix} & \frac{N(r_j)^w}{w!} & \end{pmatrix}$$

It is now easy to see that the required value of $\mathcal{I}$ is given in terms of a finite sum of elementary integrals of the type

$$\mathcal{I} = \sum_{j_0=1}^{p} \sum_{w_{j_0}=1}^{r_{j_0}-1} \cdots \sum_{j_k=1}^{p} \sum_{w_{j_k}=1}^{r_{j_k}-1} \mathcal{J}(j_0 \dots j_k; w_{j_0} \dots w_{j_k})$$

where

$$\mathcal{J}(j_0 \dots j_k; w_{j_0} \dots w_{j_k}) = \int\limits_{u_0+\cdots+u_k=1} e^{u_0\lambda_{j_0}} u_0^{w_{j_0}} M(j_0, w_{j_0}) \prod_{l=1}^{k} P_l e^{u_l\lambda_{j_l}} u_l^{w_{j_l}} M(j_l, w_{j_l}) d\mathbf{u}.$$

In the summation above each of $j_0, j_1, \cdots, j_k$ take values independently from $1, 2, \cdots, p$ and for each j_l the corresponding w_{j_l} takes values in $1, 2, \cdots, r_{j_l}$. Note that p is the number of distinct eigenvalues of D_0 and r_{j_l} denotes the multiplicity of the eigenvalue λ_{j_l}.

It can be seen that the integrating factors in $\mathcal{J}(j_0 \dots j_k; w_{j_0} \dots w_{j_l})$ are only scalars which can be grouped together such that

$$\mathcal{J}(j_0 \dots j_k; w_{j_0} \dots w_{j_l}) = M(j_0, w_{j_0}) \prod_{l=1}^{k} P_l M(j_l, w_{j_l}) \int\limits_{u_0+\cdots+u_k=1} e^{\sum_{l=0}^{k} u_l\lambda_{j_l}} \prod_{l=0}^{k} u_l^{w_{j_l}} d\mathbf{u}$$

It is now clear that we only need to evaluate the value of $\int\limits_{u_0+\cdots+u_k=1} e^{\sum_{l=0}^{k} u_l\lambda_{j_l}} \prod_{l=0}^{k} u_l^{w_{j_l}} d\mathbf{u}$. Using the result (14) we can exactly evaluate $\mathcal{I}$. Note that for implementing directly (14) in this case we need all λ_{j_l}'s distinct, otherwise, we then need to initially collapse to a single u_i all those u_l's such that λ_{j_l}'s share the same value.

4.2. Matrix exponential approach

A novel idea for calculating $\mathcal{I}$ is reported in Carbonell et al. (2007) who describe a method for calculating the matrix integrals of the type

$$\int\limits_{u_0+\cdots+u_k=t} e^{A_0 u_0} P_1 e^{A_1 u_1} \cdots P_k e^{A_k u_k} d\mathbf{u}$$

where A_i's and P_i's are square of dimension $m \times m$. Their method relies heavily on the matrix exponential function and expands to any k the approach of Van Loan (1978) for $k \leq 4$. In particular, they show that the resulting matrix above is in fact the top-right $m \times m$ sub-matrix of $\exp(t\mathcal{A})$, where $\mathcal{A}$ is a two-diagonal square matrix of dimension $(k+1)m$ defined as

$$\mathcal{A} = \begin{pmatrix} A_0 & P_1 & 0 & \cdots & 0 \\ 0 & A_1 & P_2 & 0 & \cdots 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & A_{k-1} & P_k \\ 0 & \cdots & 0 & 0 & A_k \end{pmatrix}.$$

Our expectation $\mathcal{I}$ is the top-right $m \times m$ sub-matrix of $\exp(\mathcal{A})$ where all A_i's are equal to D_0.

The implementation of this approach is straightforward, but rather inefficient since we need to calculate the $(k+1)m \times (k+1)m$ matrix $\exp(\mathcal{A})$ and extract only a small part. If its dimension however is unmanageably large for the computer we can apply the same result to low order matrices as shown below.

Applying the result of Carbonell et al. (2007) for a scalar case i.e. all $A_i = \lambda_i$ are numbers and $P_i = 1$ see that the corresponding integral in (14) for $n_i = 0$ is simply the top right entry of $\exp(\mathcal{A})$, where

$$\mathcal{A} = \begin{pmatrix} \lambda_0 & 1 & 0 & \cdots & 0 \\ 0 & \lambda_1 & 1 & 0 & \cdots 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & \lambda_{k-1} & 1 \\ 0 & \cdots & 0 & 0 & \lambda_k \end{pmatrix}.$$

One can easily show that the values in (14) for $n_i \neq 0$ can be similarly obtained by expanding each dimension of $\mathcal{A}$ to $\sum_{i=0}^{k} n_i + k + 1$ such that each λ_i is repeated $n_i + 1$ times and the resulting value obtained after evaluating $\exp(\mathcal{A})$ needs re-scaling by $\prod_{i=0}^{k} \Gamma(n_i + 1)$. We can use this method to then evaluate the elementary integrals $\mathcal{J}(j_0 \ldots j_k; w_{j_0} \ldots w_{j_l})$ in the direct approach in subsection 4.1.2.

5. Numerical examples

The purpose of this section is to show that the proposed algorithm can indeed be implemented in a tractable way on a normal PC. We utilized the first matrix exponential approach of subsection 4.2 where $\mathcal{A}$ has matrix blocks and performed the calculations in the Statistical package R.

Example 1 We consider an application where we know that all inter–arrival times have an exponential distribution. This makes an estimation simpler as we can set the off–diagonal elements of $\hat{D}_0$ as zero. The EM algorithm guarantees that initial estimates of zero remain zero, see equations

(11) and (12). We set the original parameters as

$$D_0 = \begin{pmatrix} -0.2 & 0 \\ 0 & -5 \end{pmatrix} \quad \text{and} \quad D_1 = \begin{pmatrix} 0 & 0.2 \\ 0.5 & 4.5 \end{pmatrix}$$

With these parameters we ran a simulation of 500 arrivals, yielding a time series of $N = 353$ intervals. This served as the input to our EM algorithm. The initial estimates were set as

$$\hat{D}_0^{(0)} = \begin{pmatrix} -N/d & 0 \\ 0 & -M \end{pmatrix} \quad \text{and} \quad \hat{D}_1^{(0)} = \begin{pmatrix} N/2d & N/2d \\ M/2 & M/2 \end{pmatrix}$$

where $M = \max\{z_n : n \leq N\}$ is the maximal number of arrivals within one interval, d is the total number of intervals without arrivals, and N is the total number of intervals in the time series. After 28 EM steps, the estimates for D_0 and D_1 are

$$\hat{D_0} = \begin{pmatrix} -0.207 & 0.000 \\ 0.000 & -4.727 \end{pmatrix} \quad \text{and} \quad \hat{D_1} = \begin{pmatrix} 0.000 & 0.207 \\ 0.555 & 4.172 \end{pmatrix}$$

The likelihood of the time series under these estimates is $\hat{l} = 2.017871e - 203$ as compared to the likelihood $l = 6.151147e - 204$ under the original parameters. Note that the qualitative entry $D_{0;11} = 0$ has been found by the algorithm on its own.

Example 2 The second example is a Markov–modulated Poisson process (MMPP). The original parameters were set as

$$D_0 = \begin{pmatrix} -1 & 0.5 \\ 0.5 & -2 \end{pmatrix} \quad \text{and} \quad D_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1.5 \end{pmatrix}$$

Again, we used these parameters to run a simulation of 500 arrivals, yielding a time series of $N = 479$ intervals. This served as the input to our EM algorithm. The initial estimates were set as

$$\hat{D}_0^{(0)} = \begin{pmatrix} -N/d & N/2d \\ M/2 & -M \end{pmatrix} \quad \text{and} \quad \hat{D}_1^{(0)} = \begin{pmatrix} N/2d & 0 \\ 0 & M/2 \end{pmatrix}$$

with M and d as defined in example 1. After 142 steps, the algorithm produced the following estimates for D_0 and D_1:

$$\hat{D_0} = \begin{pmatrix} -1.509 & 1.130 \\ 1.370 & -3.220 \end{pmatrix} \quad \text{and} \quad \hat{D_1} = \begin{pmatrix} 0.378 & 0.000 \\ 0.000 & 1.850 \end{pmatrix}$$

The likelihood of the time series under these estimates is $\hat{l} = 7.820401e - 285$, under the original parameters it is $l = 7.218375e - 285$.

Example 3 Now a full Markovian arrival process with two phases: Original parameters are

$$D_0 = \begin{pmatrix} -2.5 & 1 \\ 2.5 & -5 \end{pmatrix} \quad \text{and} \quad D_1 = \begin{pmatrix} 1 & 0.5 \\ 1.5 & 1 \end{pmatrix}$$

Again, we used these parameters to run a simulation of 500 arrivals, this time yielding a time series of $N = 293$ intervals. The initial estimators were set as

$$\hat{D}_0^{(0)} = \begin{pmatrix} -N/d & N/3d \\ M/3 & -M \end{pmatrix} \quad \text{and} \quad \hat{D}_1^{(0)} = \begin{pmatrix} N/3d & N/3d \\ M/3 & M/3 \end{pmatrix}$$

After only two steps, the estimates for D_0 and D_1 are

$$\hat{D_0} = \begin{pmatrix} -2.432 & 0.992 \\ 2.521 & -4.948 \end{pmatrix} \quad \text{and} \quad \hat{D_1} = \begin{pmatrix} 0.961 & 0.479 \\ 1.458 & 0.970 \end{pmatrix}$$

The likelihood of the time series under these estimates is $\hat{l} = 6.623291e - 209$, under the original parameters it is $l = 4.625924e - 209$.

If we apply the same algorithm to the first part of the same data such that there are $N = 200$ intervals only, we obtain after 28 steps

$$\hat{D_0} = \begin{pmatrix} -2.323 & 1.175 \\ 3.361 & -6.757 \end{pmatrix} \quad \text{and} \quad \hat{D_1} = \begin{pmatrix} 0.576 & 0.572 \\ 1.635 & 1.761 \end{pmatrix}$$

Here the likelihoods are $\hat{l} = 9.80752e - 147$ under the estimates and $l = 2.817786e - 147$ under the original parameters.

Example 4 The last example deals with real data taken from measurements of fetal lamb movements. They have been analysed in Leroux and Puterman (1992) via discrete time hidden Markov models. Here we apply our continuous time model to these data. In Leroux and Puterman (1992) the assumption was that in each interval the number of counts follows a Poisson distribution. The equivalent assumption in a continuous time model is that of exponential inter-arrival times. We can model this by setting D_0 to be diagonal, which means that the underlying phase can change only upon an arrival (i.e. together with an observed movement). In order to find the most suitable number m of phases, we just try increasing values of m until the likelihood gain does not appear to be worthwhile anymore.

The estimates for $m = 2$ are

$$\hat{D_0} = \begin{pmatrix} -0.243 & 0.000 \\ 0.000 & -2.775 \end{pmatrix} \quad \text{and} \quad \hat{D_1} = \begin{pmatrix} 0.222 & 0.021 \\ 0.435 & 2.340 \end{pmatrix}$$

where the achieved likelihood is $3.638315e - 78$. For $m = 3$ the estimates are

$$\hat{D_0} = \begin{pmatrix} -0.096 & 0.000 & 0.000 \\ 0.000 & -0.548 & 0.000 \\ 0.000 & 0.000 & -3.631 \end{pmatrix} \quad \text{and} \quad \hat{D_1} = \begin{pmatrix} 0.059 & 0.028 & 0.009 \\ 0.044 & 0.504 & 0.000 \\ 0.221 & 0.000 & 3.410 \end{pmatrix}$$

They generate a likelihood of $1.264017e - 73$. The estimates for $m = 4$ yield a likelihood of $1.275225e - 72$. This last likelihood gain appears as too small to justify an extra phase. Hence we stop here and decide for the model with three phases.

It is remarkable that the qualitative entries $\hat{D}_1(2,3) = \hat{D}_1(3,2) = 0$ have been picked up by the discrete time model in Leroux and Puterman (1992), table 4 under $m = 3$, too. The interpretation is that phase 1 serves as an intermediate phase, over which also changes between phases 2 and 3 need to occur.

Acknowledgement

We are obliged to Dr. Rolando Biscay from the Instituto de Cibernética, Matemática y Física in Habana, Cuba, for sending us the manuscript Carbonell et al. (2007), which contains a crucial result for the computational part of this paper. We further thank Martin Ridout from the University of Kent at Canterbury, UK, for providing the suitable data set for example 4.

References

Albert, A. (1962). Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Ann. Math. Stat. 33*, 727–753.

Asmussen, S. (1997). Phase-type distributions and related point processes: Fitting and recent advances. In: Chakravarthy and Alfa (eds.), Matrix-analytic methods in stochastic models, NY: Marcel Dekker. Lect. Notes Pure Appl. Math. 183, pp.137–149 .

Asmussen, S. and G. Koole (1993). Marked point processes as limits of Markovian arrival streams. *J. Appl. Probab. 30*(2), 365–372.

Asmussen, S., O. Nerman, and M. Olsson (1996). Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat. 23*(4), 419–441.

Bladt, M., A. Gonzalez, and S. Lauritzen (2003). The estimation of phase-type related functionals using Markov chain Monte Carlo methods. *Scandinavian Actuarial Journal 2003*(4), 280–300.

Bladt, M. and M. Soerensen (2005). Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B 67*(3), 395–410.

Breuer, L. (2002). An EM Algorithm for Batch Markovian Arrival Processes and its Comparison to a Simpler Estimation Procedure. *Annals of Operations Research 112*, 123–138.

Carbonell, F., Jimenez, J.C., and Pedroso, L.M. (2007). Computing multiple integrals involving matrix exponentials. *Instituto de Cibernética, Matemática y Física, Habana, Cuba. Manuscript*

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Discussion. *J. R. Stat. Soc., Ser. B 39*, 1–38.

Fearnhead, P. and C. Sherlock (2006). An exact Gibbs sampler for the Markov–modulated Poisson process. *J. R. Statist. Soc. B 68*(5), 767–784.

Jewell, N. P. (1982). Mixtures of exponential distributions. *Ann. Stat. 10*, 479–484.

Kent, J. (1994). The complex Bingham distribution and shape analysis. *J.R. Statist. Soc. Series B 56*, 285–289.

Kume, A. and A. Wood (2007). On the normalising constant of the Bingham distribution. *Statistics and Probability Letters*. *77*, 832–837.

Leroux, B. G. and M. L. Puterman (1992). Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models. *Biometrics 48*, 545–558.

Lucantoni, D. M. (1991). New results on the single server queue with a batch Markovian arrival process. *Commun. Stat., Stochastic Models 7*(1), 1–46.

Lucantoni, D. M. (1993). The BMAP/G/1 Queue: A Tutorial. In L. Donatiello and R. Nelson (Ed.), *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, pp. 330–358. Springer.

McLachlan, G. J. and T. Krishnan (1997). *The EM algorithm and extensions.* New York, NY: John Wiley & Sons.

Meng, X.-L. and D. van Dyk (1997). The EM algorithm - an old folk-song sung to a fast new tune. *J. R. Stat. Soc., Ser. B 59*(3), 511–567.

Neuts, M. F. (1979). A versatile Markovian point process. *J. Appl. Probab. 16*, 764–774.

Nijenhuis A. and Herbert S. W. (1978). Combinatorial Algorithms. Academic Press.

Ryden, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Comput. Stat. Data Anal. 21*(4), 431–447.

Ryden, T. (1997). Estimating the order of continuous phase-type distributions and Markov-modulated Poisson processes. *Commun. Stat., Stochastic Models 13*(3), 417–433.

Sundberg, R. (1974). Maximum Likelihood Theory for Incomplete Data from an Exponential Family. *Scand. J. Statist. 1*, 49–58.

Van Loan, C. F. (1978). Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control. 23*, 395–404.