Filtering and the EM-algorithm for the Markovian arrival process

James Ledoux

Filtering and the EM-algorithm for the Markovian Arrival Process

James Ledoux
IRMAR UMR 6625 & INSA de Rennes
20 avenue des Buttes de Coësmes
35708 Rennes Cedex 7
FRANCE
email: james.ledoux@insa-rennes.fr

4 July 2006

Abstract

In this paper, we deal with the so-called Markovian Arrival process (MAP). An MAP is thought of as a partially observed Markov process, so that the Expectation-Maximization (EM) algorithm is a natural way to estimate its parameters. Then, non-linear filters of basic statistics related to the MAP must be computed. The forward-backward principle is the basic way to do it. Here, bearing in mind a filter-based formulation of the EM-algorithm proposed by Elliott, these filters are shown to be the solution of non-linear stochastic differential equations (SDEs) which allows a recursive computation. This is well suited for processing large data sets. We also derive linear SDEs or Zakai equations for the so-called unnormalized filters.

Key-Words Hidden Markov process, Point process, Innovations method, Zakai's filter, Queuing theory, Software reliability

1 Introduction

A major issue in stochastic modeling is to calibrate the models from data. Here, we focus on a class of Markovian models known as the Markovian Arrival Processes (MAP) (e.g. see Neuts (1989), Asmussen (2000)). An MAP is formally defined as a bivariate Markov process $(N_t, X_t)_{t\geq 0}$ where $(N_t)_{t\geq 0}$ is a counting process of "arrivals", $(X_t)_{t\geq 0}$ is a Markov process with a finite state space, say $\{1, \ldots, n\}$, and the transition probabilities of $(N_t, X_t)_{t\geq 0}$ satisfy the following additivity property: for any $k = 0, 1$, $i, j = 1, \ldots, n$, $m \in \mathbb{N}$ and $0 \leq s < t$

$$\mathbb{P}\{N_t = k + m, X_t = j \mid N_s = m, X_s = i\} = \mathbb{P}\{N_t - N_s = k, X_t = j \mid X_s = i\}.$$

The other transition probabilities are zero. The property above implies that the entries of the generator A of $(N_t, X_t)_{t\geq 0}$ satisfy similar conditions: for $k = 0, 1,\ i, j = 1, \ldots, n$ and $m \in \mathbb{N}$

$$A\big((k+m, j), (m, i)\big) = A\big((k, j), (0, i)\big) := D_k(j, i)$$

and the other entries are zero. The equalities above define two $n \times n$-matrices D_0, D_1. The non-negative numbers $D_0(j, i), D_1(j, i)$ $(j \neq i)$ represent the rates at which $(X_t)_{t\geq 0}$ jumps from state i to j with no arrival and one arrival respectively. The non-negative entry $D_1(i, i)$ is the rate at which one arrival occurs, $(X_t)_{t\geq 0}$ staying in state i. Note that the Markov process $(X_t)_{t\geq 0}$ has the generator $Q = D_0 + D_1$. Listing the state space $\mathbb{N} \times \{1, \ldots, n\}$ in lexicographic order, the generator of an MAP has the form

$$A = \begin{pmatrix} D_0 & \mathbf{0} & \cdots \\ D_1 & D_0 & \ddots \\ \mathbf{0} & D_1 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}. \tag{1}$$

The so-called Markov Modulated Markov Process is the special instance of an MAP obtained in setting

$$D_1 := \mathrm{Diag}(\lambda(i)) \quad D_0 := Q - \mathrm{Diag}(\lambda(i))$$

where $\mathrm{Diag}(\lambda(i))$ is a diagonal matrix with ith diagonal entry $\lambda(i)$ and Q is the generator of the Markov process $(X_t)_{t\geq 0}$. In this model, the arrival instants constitute a Poisson process with intensity $\lambda(i)$ during a sojourn time of $(X_t)_{t\geq 0}$ in state i. The main properties of such a class of processes may be found in Fischer and Meier-Hellstern (1993), for instance.

MAPs have been introduced in queuing theory in order to consider Markovian input streams for queuing systems with non-independent inter-arrival durations. The Markov property allows to deal with analytically tractable models. This class of models has gained widespread use in stochastic modeling of communication systems, in reliability for systems and many other applications areas (e.g see Neuts (1995) for an extensive bibliography). Our motivation for dealing with MAPs originates in the software reliability modeling using an "architecture-based" approach. Indeed, a standard model in this context was provided by Littlewood (1975). It has inspired most other works (see Goseva-Popstojanova and Trivedi (2001) for a recent survey). The failure process associated with such a kind of models turns to be a point process associated with a specific MAP. Thus, in our context of software reliability modeling, $(N_t)_{t\geq 0}$ is the counting process of failures and $(X_t)_{t\geq 0}$ is interpreted to be a Markovian model of the flow of control between the modules of a software.

Although of widespread use for more than twenty years now, the statistical analysis and specifically fitting of MAPs to data, has been discussed only recently. This is a major issue, especially in software reliability modeling where, to the best of our knowledge, little has been done for the statistical estimation of the parameters of the architecture-based models. Since an MAP is specified by the matrices D_0, D_1, the fitting of MAPs to data

requires the estimation of the non-negative parameter vector

$$\theta = \{D_k(j,i), \quad k = 0, 1 \quad i, j = 1, \ldots, n\}. \tag{2}$$

The only available data is the observation of arrivals. In that perspective, the process $(N_t, X_t)_{t\geq 0}$ can be thought of as a partially observed Markov process or a hidden Markov process. The observed process is the counting process of arrivals $(N_t)_{t\geq 0}$ and the state or hidden process is the finite Markov process $(X_t)_{t\geq 0}$. The EM-algorithm is a standard way to estimate the parameters of hidden Markov processes. Specifically, it has been used by Rydén (1996) for the Markov Modulated Poisson Process, by Asmussen (1996) for the Phase-Type distributions, by Breuer (2002), Klemm et al. (2003) for general MAPs. The numerical experiments reported in their studies show that the EM-algorithm works well in general. All these works use the standard forward-backward principle which is based on data processing in batch. Due to the backward pass through the data, the storage cost is linear in the number of observations. Elliott et al. (1995) has proposed a filter-based EM-algorithm where the standard forward-backward form of the E-step of the algorithm is replaced by a single forward pass procedure. The main advantages are that "on-line" estimation is allowed and the storage cost does not depend on the number of observations so that very large data sets can be processed. To implement the filter-based approach, finite-dimensional recursive filters for various statistics related to MAPs must be found (if there exist). The aim of this paper is to provide such finite-dimensional filters.

The paper is organized as follows. In Subsection 2.1, the basic material on stochastic calculus needed throughout the paper is introduced. Next, the EM-algorithm and the forward-backward strategy are discussed for MAPs. In Subsection 2.3, the filters associated with the statistics of interest for using the EM-algorithm are shown to be the solutions of non-linear stochastic differential equations (SDEs). In Subsection 2.4, the so-called unnormalized/Zakai filters are proved to be the solution of linear SDEs. These SDEs allow a recursive computation of filters involved in the re-estimation formulas of the parameters in (2). They are the main contributions of the paper. The forward-backward and filter-based strategies are briefly compared in Subsection 2.5. Concluding comments are reported in Section 3.

2 Finite-dimensional filters

Main notation and convention

- Vectors are column vectors. Row vectors are obtained by means of the transpose operator $(\cdot)^\top$. The ith component of any vector v is denoted by $v(i)$.

  For any pair of vectors $u, v \in \mathbb{R}^n$, $\langle u, v\rangle = u^\top v$, is the usual scalar product in $\mathbb{R}^n$.

  $\mathbf{1}$ is a n-dimensional vector with each entry equals to one.

- For any right-continuous with left-hand limits (rcll) function $t \mapsto f_t$, the left-hand limit at t of f is denoted by f_{t-} and $\Delta f_t := f_t - f_{t-}$ for $t > 0$ is the jump of the function at time t. We set $\Delta f_0 := f_0$.

- The state space of the Markov process $(X_t)_{t\geq 0}$ is assumed to be $\mathscr{X} := \{e_i, i = 1, \ldots, n\}$, where e_i is the ith vector of the canonical basis of $\mathbb{R}^n$.
  With this convention, the indicator function $1_{\{X_t=e_i\}}$ of the set $\{X_t = e_i\}$ is the ith component of vector X_t, that is $\langle X_t, e_i\rangle$. In other words, for any time t, the vector X_t is the n-dimensional vector

$$X_t = \left(1_{\{X_t=e_i\}}\right)_{i=1}^n.$$

  Therefore, the sum of the components of vector X_t, $\langle \mathbf{1}, X_t\rangle$, is equal to 1 for any time t.

- All processes are assumed to be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The internal filtrations of processes $(N_t)_{t\geq 0}$ and $(N_t, X_t)_{t\geq 0}$ are denoted by $\mathbb{F}^N := (\mathbb{F}_t^N)_{t\geq 0}$ and $\mathbb{F} := (\mathbb{F}_t)_{t\geq 0}$ respectively, where $\mathbb{F}_t^N := \sigma(N_s,\ s \leq t)$ and $\mathbb{F}_t := \sigma(N_s, X_s\ s \leq t)$. These filtrations are assumed to be complete, that is $\mathbb{F}_0, \mathbb{F}_0^N$ contain all the sets of $\mathbb{P}$-probability zero of $\mathcal{F}$.

- For any integrable $\mathbb{F}$-adapted random process $(Z_t)_{t\geq 0}$, the conditional expectation $\mathbb{E}[Z_t \mid \mathbb{F}_t^N]$ is denoted by $\widehat{Z}_t$ and $(\widehat{Z}_t)_{t\geq 0}$ is called the *filter* associated with the process $(Z_t)_{t\geq 0}$.

2.1 Basic material on the observed/hidden processes

In this paper, we are concerned with the following basic statistics of an MAP

$$\begin{aligned} & \mathcal{L}_t^{1,ji} := \sum_{0<s\leq t} \Delta N_s \langle X_s, e_j\rangle\langle X_{s-}, e_i\rangle = \int_0^t \langle X_s, e_j\rangle\langle X_{s-}, e_i\rangle \, \mathrm{d}N_s \\ j \neq i \quad & \mathcal{L}_t^{0,ji} := \sum_{0<s\leq t} (1 - \Delta N_s)\langle X_s, e_j\rangle\langle X_{s-}, e_i\rangle \\ & \mathcal{O}_t^{(i)} := \int_0^t \langle X_{s-}, e_i\rangle \, \mathrm{d}s. \end{aligned} \tag{3}$$

For $i \neq j$, the first – second – statistic is *the number of jumps of* $(X_t)_{t\geq 0}$ *from state* e_i *to state* e_j *coming with one – no – jump of the counting process over the interval* $]0, t]$. $\mathcal{L}_t^{1,ji}$ can also be thought of as the number of arrivals coming with a jump of $(X_t)_{t\geq 0}$ from state e_i to e_j over the interval $]0, t]$. $\mathcal{L}_t^{1,ii}$ is the number of arrivals up to time t, $(X_t)_{t\geq 0}$ staying in state e_i at each arrival instant. The third statistic is the *sojourn time of* $(X_t)_{t\geq 0}$ *in the state* e_i in the interval $[0, t]$. This family of statistics is introduced because their filters appear in the re-estimation formulas (10) used for estimating the parameters of an MAP with the EM-algorithm.

The basic material below can be found in Bremaud (1981), Klebaner (1998) for instance. We report here a $\mathbb{F}$-semi-martingale (or Doob-Meyer here) representation of the counting processes $(N_t)_{t\geq 0}, (\mathcal{L}_t^{0,ji})_{t\geq 0}, (\mathcal{L}_t^{1,ji})_{t\geq 0}$. It is clear from their definition that $(N_t)_{t\geq 0}, (\mathcal{L}_t^{0,ji})_{t\geq 0}, (\mathcal{L}_t^{1,ji})_{t\geq 0}$ are counters of specific transitions in $(N_t, X_t)_{t\geq 0}$. Now, the

$\mathbb{F}$-semi-martingale decomposition of the number of transitions $N_t(y,x)$ of $(N_t, X_t)_{t\geq 0}$ from state x to state y at time t is known to be

$$N_t(y,x) = \int_0^t A(y,x)\, 1_{\{(N_{s-},X_{s-})=x\}}\, \mathrm{d}s + M_t(y,x)$$

where $(M_t(y,x))_{t\geq 0}$ is a $\mathbb{F}$-*martingale* and A is the generator of $(N_t, X_t)_{t\geq 0}$. Then, it is easily deduced from the special structure of A (see (1)) that the $\mathbb{F}$-semi-martingale decomposition of the counting processes above are

$$N_t = \int_0^t \lambda_s\, \mathrm{d}s + \mathcal{M}_t \quad \text{with } \lambda_s := \langle \mathbf{1}, D_1 X_{s-}\rangle \tag{4}$$

$$\mathcal{L}_t^{k,ji} = \int_0^t D_k(j,i)\langle X_{s-}, e_i\rangle\, \mathrm{d}s + \mathcal{M}_t^{\mathcal{L}^{k,ji}} \quad k = 0,1 \tag{5}$$

where $j \neq i$ for $k=0$, and $(\mathcal{M})_{t\geq 0}, (\mathcal{M}^{\mathcal{L}^{k,ji}})_{t\geq 0}$ are $\mathbb{F}$-martingales. The process $(\lambda_t)_{t>0}$ is the so-called *stochastic intensity* of $(N_t)_{t\geq 0}$with respect to the filtration $\mathbb{F}$.

The basic $\mathbb{F}$-semi-martingale decomposition of the Markov process $(X_t)_{t\geq 0}$ is

$$X_t = X_0 + \int_0^t QX_{s-}\, \mathrm{d}s + \mathcal{M}_t^X \tag{6}$$

where $(\mathcal{M}_t^X)_{t\geq 0}$ is a $\mathbb{F}$-martingale (and a martingale with respect to the internal filtration of $(X_t)_{t\geq 0}$ as well). We recognize in (4) and (6) a standard representation of a continuous-time hidden Markov process, with $(X_t)_{t\geq 0}$ as state process and $(N_t)_{t\geq 0}$ as observed process. The observation and state "noises" $(\mathcal{M}_t)_{t\geq 0}, (\mathcal{M}_t^X)_{t\geq 0}$ are correlated here.

The following results will be used throughout the proofs.

(R1) The stochastic integrals in this paper are Lebesgue-Stieltjes integrals. We report here the product rule for two rcll processes $(Z_t^{(1)})_{t\geq 0}, (Z_t^{(2)})_{t\geq 0}$ having finite variations on the bounded intervals, *i.e.* having locally finite variations,

$$\begin{aligned} Z_t^{(1)} Z_t^{(2)} &= Z_0^{(1)} Z_0^{(2)} + \int_0^t Z_{s-}^{(1)} \mathrm{d}Z_s^{(2)} + \int_0^t Z_{s-}^{(1)} \mathrm{d}Z_s^{(2)} + \sum_{0<s\leq t} \Delta Z_s^{(1)} \Delta Z_s^{(2)} \\ &= Z_0^{(1)} Z_0^{(2)} + \int_0^t Z_{s-}^{(1)} \mathrm{d}Z_s^{(2)} + \int_0^t Z_s^{(1)} \mathrm{d}Z_s^{(2)}. \end{aligned}$$

Since any rcll process $(Z_t)_{t\geq 0}$ has finitely many jumps on finite intervals for almost all $\omega \in \Omega$, the following standard Lebesgue integrals

$$\int_0^t Z_s \mathrm{d}s \quad \text{or} \quad \int_0^t Z_{s-} \mathrm{d}s$$

are interchangeably used.

(R2) For any integrable process $(Z_t)_{t\geq 0}$ such that $\int_0^t \mathbb{E}\big[|Z_s|\big]\,\mathrm{d}s < +\infty$, we have that

$$\widehat{\int_0^t Z_s\,\mathrm{d}s} - \int_0^t \widehat{Z}_s\,\mathrm{d}s$$

defines a $\mathbb{F}^N$-martingale.

(R3) Let $(H_t)_{t\geq 0}$ be a $\mathbb{F}$-predictable process and $(\mathcal{M}_t)_{t\geq 0}$ be a $\mathbb{F}$-martingale of integrable variation on any bounded intervals. If $\int_0^t \mathbb{E}\big[|Z_s|\,|\mathrm{d}\mathcal{M}s|\big] < +\infty$ for any t, then $(\int_0^t H_s \mathrm{d}\mathcal{M}_s)_{t\geq 0}$ is a $\mathbb{F}$-martingale. Note that any left-continuous $\mathbb{F}$-adapted process is $\mathbb{F}$-predictable. The same statement holds replacing everywhere the filtration $\mathbb{F}$ by $\mathbb{F}^N$.

(R4) For any $\mathbb{F}$-martingale $(\mathcal{M}_t)_{t\geq 0}$, $(\widehat{\mathcal{M}}_t)_{t\geq 0}$ is a $\mathbb{F}^N$-martingale.

2.2 The EM-algorithm and the forward-backward strategy

We briefly describe the EM-algorithm for our continuous-time hidden Markov model. We refer to Breuer (2002), Klemm et al. (2003) for full details. For a fixed parameter vector θ as defined by (2), the underlying probability measure and the associated expectation are denoted by $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ respectively. X_0 or its probability distribution x_0 is assumed to be known. The observed data are supposed to be the arrival times $\{t_0, \ldots, t_K\}$, where we set $t_0 := 0$. Without loss of generality, we assume in this subsection that $t := t_K$. The likelihood of the observed data is under $\mathbb{P}_\theta$

$$l(\theta, N) := \mathbf{1}^\top \left(\prod_{l=K}^{1} D_1 \exp\big(D_0(t_l - t_{l-1})\big) \right) x_0 \tag{7}$$

and is called the *observed/incomplete data likelihood*. Now, suppose that the complete data $\{N_s, X_s, s \leq t\}$ are available. Then, the *complete data likelihood* function is, under $\mathbb{P}_\theta$,

$$L(\theta; N, X) := \prod_{i,j=1,}^{n} D_1(j,i)^{\mathcal{L}_t^{1,ji}} \prod_{i,j=1, j\neq i}^{n} D_0(j,i)^{\mathcal{L}_t^{0,ji}} \prod_{i=1}^{n} e^{D_0(i,i)\mathcal{O}_t^{(i)}} \prod_{i=1}^{n} \langle x_0, e_i\rangle^{\langle X_0, e_i\rangle}. \tag{8}$$

It can be shown that a new estimate $\widetilde{\theta} := \{\widetilde{D}_k(j,i),\ i,j = 1,\ldots,n;\ k = 0,1\}$ satisfying $l(\widetilde{\theta}, N) \geq l(\theta, N)$, is obtained as a result of the two following steps.

1. **E-step.**

   Compute the so-called *pseudo-log-likelihood* $Q(\cdot \mid \theta)$ defined by

$$Q(\theta^* \mid \theta) := \mathbb{E}_\theta\big[\log L(\theta^*; N, X) \mid \mathbb{F}_t^N\big]$$

with $\theta^* := \{D_k^*(j,i),\ i,j=1,\ldots,n;\ k=0,1\}$. It is easily seen from (8) that

$$Q(\theta^* \mid \theta) = \sum_{i,j=1,j\neq i}^{n} \log D_0^*(j,i)\ \widehat{\mathcal{L}}_t^{0,ji} + \sum_{i,j=1}^{n} \log D_1^*(j,i)\widehat{\mathcal{L}}_t^{1,ji} + \sum_{i=1}^{n} D_0^*(i,i)\widehat{\mathcal{O}^{(i)}}_t + K \quad (9)$$

where K is a constant that does not depend on the parameters, and

$$k=0,1,\ \widehat{\mathcal{L}}_t^{k,ji} := \mathbb{E}_\theta[\mathcal{L}_t^{k,ji} \mid \mathbb{F}_t^N], \quad \widehat{\mathcal{O}^{(i)}}_t := \mathbb{E}_\theta[\mathcal{O}_t^{(i)} \mid \mathbb{F}_t^N]$$

are the filters associated with the statistics defined in (3).

2. **M-step.**

   Determine $\widetilde{\theta}$ maximizing the function in (9) under the constraints that $\sum_{j=1}^{n}(D_0^*(j,i)+D_1^*(j,i)=0, i=1,\ldots n$. Using the Lagrange multipliers method, it is shown that, for $i,j=1,\ldots,n$,

$$\widetilde{D}_1(j,i) = \frac{\widehat{\mathcal{L}}_t^{1,ji}}{\widehat{\mathcal{O}^{(i)}}_t} \quad \text{and} \quad \widetilde{D}_0(j,i) = \frac{\widehat{\mathcal{L}}_t^{0,ji}}{\widehat{\mathcal{O}^{(i)}}_t} \quad \text{with } i \neq j. \quad (10)$$

Therefore, the idea is to pick up an initial value $\theta^{(0)}$ for the parameter and to iterate 1-2 as long as a stopping criterion is not satisfied. As a result, we obtain a sequence $(\theta^{(m)})_{m\geq 0}$ of estimates corresponding to non-decreasing values of the observed likelihood function (with equality iff $\theta^{(m+1)} = \theta^{(m)}$ under mild conditions). Note that the zero entries of D_ks are preserved by the procedure above.

Remark 1 We mention that formulas (10) are intuitively supported by the fact that, canceling the conditional expectation operation, we retrieve the estimators that we would obtained applying the standard Maximum Likelihood method to the complete data likelihood (8), that is if the complete data were observed.

Remark 2 In software reliability context, a priori estimates for θ using procedures reported in Goseva-Popstojanova and Trivedi (2001) can be obtained. They are based on data collected at earlier phases of the software life cycle (validation phases, integration tests,...). These estimates might appear to be rough estimates especially when the software is in operation. This motivates a re-estimation of the parameters by EM when failure data are collected during the operational life of the software.

Remark 3 In this paper, the number of hidden states n is assumed to be known. This is adequate, for example, in speech recognition where the hidden states are the elements of a finite alphabet, in architecture-based software reliability modeling where the hidden states are the modules of a piece of software. However, in many applications, this is not the case. A generic situation is when a partially observed model is used as a statistical model for fitting to empirical time series. Estimation of the number of hidden states is known to be a hard problem. It is not intended to address this fundamental issue here. In

the context of hidden Markov chains, the so-called "order estimation problem" is surveyed in (Ephraim and Merhav 2002, Section VIII). Recent progress in this direction is reported in Boucheron and Gassiat (2005). We refer the reader to these papers and the references therein for details.

This iterative procedure requires the computation of the conditional expectations in (10). For hidden Markov models, the "*forward-backward*" *formulation* of the EM-algorithm – also referred to as the Baum-Welch formulation – is the standard way to do it. This is what is done in the previously mentioned works. The basic idea is to write these conditional expectations using Fubini's theorem as

$$\widehat{\mathcal{L}}_t^{1,ji} = \int_0^t \mathbb{P}\{\Delta N_s = 1, X_{s-} = e_i, X_s = e_j \mid \mathbb{F}_t^N\}\mathrm{d}s \quad \widehat{\mathcal{O}^{(i)}}_t = \int_0^t \mathbb{P}\{X_s = e_i \mid \mathbb{F}_t^N\}\mathrm{d}s$$
$$\widehat{\mathcal{L}}_t^{0,ji} = \int_0^t \mathbb{P}\{\Delta N_s = 0, X_{s-} = e_i, X_s = e_j \mid \mathbb{F}_t^N\}\mathrm{d}s$$

so that the conditional probabilities under the integral sign must be evaluated. The probabilities $\mathbb{P}\{X_s = e_i \mid \mathbb{F}_t^N\}, s < t$ are known as the *state smoothers* and are computed according the "forward-backward" strategy. The other conditional probabilities are deduced from these state smoothers. The final form of the EM-algorithm is given by Figure 1 (see Klemm et al. (2003)).

Bearing in mind the filter-based approach pioneered by Elliott et al. (1995), it is expected that the conditional expectations in (10) are solution of finite-dimensional recursive equations, that is *finite-dimensional filters* exist. The main contribution of the paper is to provide such finite-dimensional filters in Theorem 4 and finite-dimensional unnormalized/Zakai filters in Theorem 6. It is clear from the second and third lines of the algorithm in Figure 1 that a forward and a backward pass through the data are required. In contrast, a single pass – a forward pass – through the data set is needed for the filter-based approach. We get back to the comparison of the two strategies in Subsection 2.5.

2.3 Stochastic differentials equations for the filters

In fact, we show that the filters defined, for every t, by

$$\widehat{\mathcal{O}^{(i)}X}_t := \mathbb{E}\big[\mathcal{O}_t^{(i)}X_t \mid \mathbb{F}_t^N\big] \text{ and } \widehat{\mathcal{L}^{k,ji}X}_t := \mathbb{E}\big[\mathcal{L}_t^{k,ji}X_t \mid \mathbb{F}_t^N\big], k = 0, 1$$

turn to be finite-dimensional. Then, since the sum of the components of vector X_t is equal to 1 for any t, the filters that we are interested in, are obtained as follows

$$\widehat{\mathcal{O}^{(i)}}_t = \langle \mathbf{1}, \widehat{\mathcal{O}^{(i)}X}_t\rangle, \quad \text{and} \quad \widehat{\mathcal{L}}_t^{k,ji} = \langle \mathbf{1}, \widehat{\mathcal{L}^{k,ji}X}_t\rangle.$$

We know from (4) that the $\mathbb{F}$-stochastic intensity of the counting process $(N_t)_{t\geq 0}$ is given, for $t > 0$ by $\lambda_t = \langle \mathbf{1}, D_1 X_{t-}\rangle$. Then, it can be deduced from Bremaud (1981) that the $\mathbb{F}^N$-stochastic intensity $(\widehat{\lambda}_t)_{t>0}$ of $(N_t)_{t\geq 0}$ is, for $t > 0$,

$$\widehat{\lambda}_t := \langle \mathbf{1}, D_1 \widehat{X}_{t-}\rangle, \tag{11}$$

$f_0(x) := \exp(D_0 x)$ and $f_1(x) := D_1 \exp(D_0 x)$; for $l = 1, \ldots, K$, $\Delta t_l := t_l - t_{l-1}$ with $t_0 := 0$.

Forward. $\alpha_0 := x_0$, $c_0 := 1$, for $l = 1, \ldots, K, \alpha_l := f_1(\Delta t_l)\alpha_{l-1}$, $c_l := \mathbf{1}^\top \alpha_l$

Backward. $\beta_{K+1}^\top := \mathbf{1}^\top$, and for $l = K, \ldots, 1$ $\beta_l^\top := \beta_{l+1}^\top f_1(\Delta t_l)$

For $i, j = 1, \ldots, n$:

$L_0^{0,ji} := 0, L_0^{1,ji} := 0$ and for $l = 1, \ldots, K$:

$$\begin{aligned} L_l^{0,ji} &:= L_{l-1}^{0,ji} + \beta_{l+1}^\top \int_{t_{l-1}}^{t_l} f_1(t_l - s)\, e_j\ D_0(j,i)\ e_i^\top f_0(s - t_{l-1})\, \mathrm{d}s\ \alpha_{l-1} \\ L_l^{1,ji} &:= L_{l-1}^{1,ji} + \beta_{l+1}^\top e_j\ D_1(j,i)\ e_i^\top f_0(\Delta t_l)\alpha_{l-1} \\ O_l^{(i)} &:= O_{l-1}^{(i)} + \beta_{l+1}^\top \int_{t_{l-1}}^{t_l} f_1(t_l - s) e_j\, e_i^\top f_0(s - t_{l-1})\, \mathrm{d}s\ \alpha_{l-1} \end{aligned}$$

$$\widehat{\mathcal{O}^{(i)}}_{t_K} = \frac{O_K^{(i)}}{c_K} \quad \widehat{\mathcal{L}}_{t_K}^{0,ji} = \frac{\mathcal{L}_K^{0,ji}}{c_K} \quad \widehat{\mathcal{L}}_{t_K}^{1,ji} = \frac{L_K^{1,ji}}{c_K}$$

Comment. Once normalized to 1, the forward quantities $\alpha_l, l = 0, \ldots, K$ give the state filter at times $t_0, \ldots, t_K$

$$\widehat{X}_{t_l} = \mathbb{E}[X_{t_l} \mid \mathbb{F}_{t_l}^N] = (\mathbb{P}\{X_{t_l} = e_i \mid \mathbb{F}_{t_l}^N\})_{i=1}^n = \frac{\alpha_l}{c_l} \quad l = 0, \ldots, K.$$

Comment. The conditional probability $\mathbb{P}\{X_{t_l} = e_i \mid \mathbb{F}_{t_K}^N\}$ for $l = 0, \ldots, K-1$ are given by

$$\mathbb{P}\{X_{t_l} = e_i \mid \mathbb{F}_{t_K}^N\} = \frac{\beta_l^\top e_i\ e_i^\top \alpha_l}{\sum_{j=1}^n \beta_l^\top e_j\ e_j^\top \alpha_l}$$

Figure 1: Forward-Backward implementation of the filters

that is the process defined for $t \geq 0$ by

$$\mathcal{N}_t := N_t - \int_0^t \widehat{\lambda}_s \, \mathrm{d}s \tag{12}$$

is a $\mathbb{F}^N$-martingale. $(\mathcal{N}_t)_{t\geq 0}$ is called the *innovations martingale*.

Theorem 4 *Let us consider the $\mathbb{F}^N$-stochastic intensity $(\widehat{\lambda}_t)_{t>0}$ of $(N_t)_{t\geq 0}$ and the innovations martingale $(\mathcal{N}_t)_{t\geq 0}$ defined in (11) and (12) respectively.*

1. ***Filter for the state.*** *We have for any $t \geq 0$*

$$\widehat{X}_t = \widehat{X}_0 + \int_0^t Q\widehat{X}_{s-} \, \mathrm{d}s + \int_0^t \frac{D_1\widehat{X}_{s-} - \widehat{X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s} \, \mathrm{d}\mathcal{N}_s. \tag{13a}$$

2. ***Filter for the sojourn time in e_i.*** *We have for any $t \geq 0$*

$$\widehat{\mathcal{O}^{(i)}X}_t = \int_0^t \left[Q\widehat{\mathcal{O}^{(i)}X}_{s-} + \langle \widehat{X}_{s-}, e_i\rangle \, \mathrm{d}s \, e_i\right] \mathrm{d}s + \int_0^t \frac{D_1\widehat{\mathcal{O}^{(i)}X}_{s-} - \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s} \, \mathrm{d}\mathcal{N}_s. \tag{13b}$$

3. ***Filter for the numbers of specific jumps of the MAP.*** *We have for any $t \geq 0$*

$$\widehat{\mathcal{L}^{0,ji}X}_t = \int_0^t [Q\widehat{\mathcal{L}^{0,ji}X}_{s-} + D_0(j,i)\langle \widehat{X}_{s-}, e_i\rangle e_j]\mathrm{d}s + \int_0^t \frac{D_1\widehat{\mathcal{L}^{0,ji}X}_{s-} - \widehat{\mathcal{L}^{0,ji}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s} \, \mathrm{d}\mathcal{N}_s \tag{13c}$$

$$\begin{aligned}\widehat{\mathcal{L}^{1,ji}X}_t = &\int_0^t \left[Q\widehat{\mathcal{L}^{1,ji}X}_{s-} + D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle e_j\right] \mathrm{d}s \\ &+ \int_0^t \frac{D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle \, e_j + D_1\widehat{\mathcal{L}^{1,ji}X}_{s-} - \widehat{\mathcal{L}^{1,ji}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s} \, \mathrm{d}\mathcal{N}_s.\end{aligned} \tag{13d}$$

We mention that equation (13a) turns to be one of the main ingredient of the proof in Gravereaux and Ledoux (2004) of the (optimal) convergence rate of an MAP to a Poisson process when the arrivals are rare. We see that the stochastic differential equations in Theorem 4 are non-linear, so that the numerical procedure used for solving these equations has to be carefully implemented. For instance, we recall that negative probabilities may be obtained as a result of the numerical computation of the state filter of a Markov process observed in an additive Brownian noise – also called Wonham's filter – with an Euler-Maruyama discrete-time approximation (see (Kloeden et al. 1994, Chap 6)). The purpose of the next subsection is to derive linear SDEs from which the filters may be obtained. Such linear SDEs are more suited to numerical computation.

Remark 5 A filter for the number of transitions $N_t^{X,ji}$ of $(X_t)_{t\geq 0}$ from state e_i to e_j $(i \neq j)$ up to time t can be derived as those of Theorem 4. Indeed, we obtain an

SDE for $\widehat{N^{X,ji}X}_t := \mathbb{E}\big[N_t^{X,ji}X_t \mid \mathbb{F}_t^N\big]$ from Theorem 4 using the fact that $\widehat{N^{X,ji}X}_t = \widehat{\mathcal{L}^{0,ji}X}_t + \widehat{\mathcal{L}^{1,ji}X}_t$ for $i \neq j$. Then, the sum of he components of the vector $\widehat{N^{X,ji}X}_t$ gives the filter associated with $N_t^{X,ji}$. Since $D_0 + D_1 = Q$, the transition rates of $(X_t)_{t\geq 0}$ can be estimated by the EM-algorithm using

$$\widetilde{Q}_1(j,i) = \frac{\widehat{N^{X,ji}}_t}{\widehat{\mathcal{O}^{(i)}}_t}$$

as re-estimation formula.

Proof. A proof of (13a) may be found in Gravereaux and Ledoux (2004). In the sequel, $(\mathcal{M}_t)_{t\geq 0}$ – $(\widehat{\mathcal{M}}_t)_{t\geq 0}$ – will denote a generic $\mathbb{F}$-martingale – $\mathbb{F}^N$-martingale. The proof of (13b) is as follows. The product rule (R1) with the fact that $(\mathcal{O}_t^{(i)})_{t\geq 0}$ has no jump give

$$\begin{aligned} \mathcal{O}_t^{(i)}X_t &= \int_0^t \mathcal{O}_{s-}^{(i)}\,\mathrm{d}X_s + \int_0^t X_{s-}\,\mathrm{d}\mathcal{O}_s^{(i)} && (14)\\ &= \int_0^t Q\mathcal{O}_{s-}^{(i)}X_{s-}\,\mathrm{d}s + \int_0^t \langle X_{s-}, e_i\rangle e_i\,\mathrm{d}s + \mathcal{M}_t \quad \text{from (6),(3) and (R3).} && (15)\end{aligned}$$

Taking the conditional expectation with respect to $\mathbb{F}_t^N$ on each side of the equation above, and using (R2) and (R4), we obtain

$$\widehat{\mathcal{O}^{(i)}X}_t = \int_0^t Q\widehat{\mathcal{O}^{(i)}X}_{s-}\,\mathrm{d}s + \int_0^t \langle \widehat{X}_{s-}, e_i\rangle\, e_i\,\mathrm{d}s + \widehat{\mathcal{M}}_t. \tag{16}$$

The integral representation of $\mathbb{F}^N$-martingales says us that (e.g. see Bremaud (1981))

$$\widehat{\mathcal{M}}_t = \int_0^t G_s^{(i)}\,\mathrm{d}\mathcal{N}_s \tag{17}$$

where $(G_t^{(i)})_{t\geq 0}$ is a $\mathbb{F}^N$-predictable process which is called the *innovations gain*. Thus, the proof will be complete if we show that

$$G_s^{(i)} = \frac{D_1\widehat{\mathcal{O}^{(i)}X}_{s-} - \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s}. \tag{18}$$

The product $N_t \times \widehat{\mathcal{O}^{(i)}X}_t$ has the following form from the product rule (R1)

$$N_t\widehat{\mathcal{O}^{(i)}X}_t = \int_0^t N_{s-}\,\mathrm{d}\widehat{\mathcal{O}^{(i)}X}_s + \int_0^t \widehat{\mathcal{O}^{(i)}X}_{s-}\,\mathrm{d}N_s + \sum_{0<s\leq t}\Delta N_s\Delta\widehat{\mathcal{O}^{(i)}X}_s.$$

Since $\Delta\widehat{\mathcal{O}^{(i)}X}_s = G_s^{(i)}\Delta\mathcal{N}_s = G_s^{(i)}\Delta N_s$ from (16),(17),(12) and $(\Delta N_s)^2 = \Delta N_s$, the last term in the right-hand side of the inequality above is $\int_0^t G_s^{(i)}\mathrm{d}N_s$. Then, we deduce from

(12),(16) and (R3) that

$$\begin{aligned} N_t\widehat{\mathcal{O}^{(i)}X}_t &= \int_0^t N_{s-}[Q\widehat{\mathcal{O}^{(i)}X}_{s-} + \langle \widehat{X}_{s-}, e_i\rangle\, e_i]\,\mathrm{d}s + \widehat{\mathcal{M}}_t \\ &\quad + \int_0^t \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{\lambda}_s\,\mathrm{d}s + \int_0^t G^{(i)}_s\widehat{\lambda}_s\,\mathrm{d}s + \widehat{\mathcal{M}}_t. \end{aligned} \tag{19}$$

Next, the product $N_t \times \mathcal{O}^{(i)}_t X_t$ is rewritten from the rule product (R1) and the fact that $(\mathcal{O}^{(i)}_t)_{t\geq 0}$ has no jump as

$$\begin{aligned} N_t\mathcal{O}^{(i)}_t X_t &= \int_0^t N_{s-}\,\mathrm{d}(\mathcal{O}^{(i)}X)_s + \int_0^t \mathcal{O}^{(i)}_{s-}X_s\,\mathrm{d}N_s \\ &= \int_0^t N_{s-}[Q\mathcal{O}^{(i)}_{s-}X_{s-} + \langle \widehat{X}_{s-}, e_i\rangle\, e_i]\,\mathrm{d}s + \int_0^t \mathcal{O}^{(i)}_{s-}X_s\,\mathrm{d}N_s + \mathcal{M}_t \end{aligned}$$

from (15) and (R3). Let us compute the last integral in the right-hand side of equality above. Since $\sum_j \langle X_s, e_j\rangle = \sum_k \langle X_{s-}, e_k\rangle = 1$, and $\Delta N_s \langle X_s, e_j\rangle\langle X_{s-}, e_k\rangle = \Delta\mathcal{L}^{1,jk}_s$ we find that

$$\begin{aligned} \int_0^t \mathcal{O}^{(i)}_{s-}X_s\,\mathrm{d}N_s &= \sum_{0<s\leq t} \mathcal{O}^{(i)}_{s-}X_s\Delta N_s = \sum_j e_j \sum_k \int_0^t \mathcal{O}^{(i)}_{s-}\,\mathrm{d}\mathcal{L}^{1,jk}_s \\ &= \int_0^t \mathcal{O}^{(i)}_{s-}\sum_j e_j \sum_k D_1(j,k)\langle X_{s-}, e_k\rangle\,\mathrm{d}s + \mathcal{M}_t \quad \text{from (5) and (R3)} \\ &= \int_0^t \mathcal{O}^{(i)}_{s-}D_1X_{s-}\,\mathrm{d}s + \mathcal{M}_t. \end{aligned}$$

Then, we deduce from the last equality that

$$N_t\mathcal{O}^{(i)}_t X_t = \int_0^t N_{s-}[Q\mathcal{O}^{(i)}_{s-}X_{s-} + \langle X_{s-}, e_i\rangle\, e_i]\,\mathrm{d}s + \int_0^t D_1\mathcal{O}^{(i)}_{s-}X_{s-}\,\mathrm{d}s + \mathcal{M}_t.$$

Taking the conditional expectation on both sides of the previous formula and using (R2) and (R4), a second decomposition of the semi-martingale $(N_t\widehat{\mathcal{O}^{(i)}X}_t)_{t\geq 0}$ is

$$N_t\widehat{\mathcal{O}^{(i)}X}_t = \int_0^t N_{s-}[Q\widehat{\mathcal{O}^{(i)}X}_{s-} + \langle \widehat{X}_{s-}, e_i\rangle\, e_i]\,\mathrm{d}s + \int_0^t D_1\widehat{\mathcal{O}^{(i)}X}_{s-}\,\mathrm{d}s + \widehat{\mathcal{M}}_t. \tag{20}$$

We know that the locally finite variations part of the decomposition of a special semi-martingale is unique. Then, we identify the corresponding terms in the decompositions (19) and (20), that is the Lebesgue integrals. The expression (18) of the gain $G^{(i)}$ follows easily.

The formulas (13c,13d) are shown in the same way. We only provide the main steps of the computation for $(\widehat{\mathcal{L}^{1,ji}X}_t)_{t\geq 0}$. The product rule (R1) and formulas (4)–(6) allow us to write

$$\mathcal{L}^{1,ji}_t X_t = \int_0^t [Q\mathcal{L}^{1,ji}_{s-}X_{s-}\,\mathrm{d}s + D_1(j,i)\langle X_{s-}, e_i\rangle\, e_j]\,\mathrm{d}s + \mathcal{M}_t.$$

Taking the conditional expectation with respect to $\mathbb{F}_t^N$ on both sides of the equations above, using (R2),(R4) and the representation theorem of the $\mathbb{F}^N$-martingales, we obtain

$$\widehat{\mathcal{L}^{1,ji}X}_t = \int_0^t [Q\widehat{\mathcal{L}^{1,ji}X}_{s-} + D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle\, e_j]\,\mathrm{d}s + \int_0^t G_s^{(\mathcal{L})}(\mathrm{d}N_s - \widehat{\lambda}_s\,\mathrm{d}s) \quad (21)$$

where $G^{(\mathcal{L})}$ is the innovations gain.

A first $\mathbb{F}^N$-representation of the semi-martingale $N_t\widehat{\mathcal{L}^{1,ji}X}_t$ is obtained from (R1),(21), (4) and (R3)

$$\begin{aligned} N_t\widehat{\mathcal{L}^{1,ji}X}_t = &\int_0^t N_{s-}[Q\widehat{\mathcal{L}^{1,ji}X}_{s-} + D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle\, e_j]\,\mathrm{d}s + \int_0^t \widehat{\mathcal{L}^{1,ji}X}_{s-}\widehat{\lambda}_s\,\mathrm{d}s \\ &+ \int_0^t G_s^{(\mathcal{L})}\widehat{\lambda}_s\,\mathrm{d}s + \widehat{\mathcal{M}}_t. \end{aligned} \quad (22)$$

Next, the product $N_t \times (\mathcal{L}_t^{1,ji}X_t)$ may be rewritten using (R1) as

$$\begin{aligned} N_t\mathcal{L}_t^{1,ji}X_t = &\int_0^t N_{s-}[Q\mathcal{L}_{s-}^{1,ji}X_{s-} + D_1(j,i)\langle X_{s-}, e_i\rangle\, e_j]\,\mathrm{d}s \\ &+ \int_0^t D_1(j,i)\langle X_{s-}, e_i\rangle\,\mathrm{d}s\, e_j + \int_0^t D_1\mathcal{L}_{s-}^{1,ji}X_{s-}\,\mathrm{d}s + \mathcal{M}_t. \end{aligned}$$

Conditioning with respect to $\mathbb{F}_t^N$ on both sides of the previous formula and using (R2) and (R4) lead to a second decomposition of the special semi-martingale $(N_t \times \mathcal{L}_t^{1,ji}X_t)_{t\geq 0}$

$$\begin{aligned} N_t\widehat{\mathcal{L}^{1,ji}X}_t = &\int_0^t N_{s-}[Q\widehat{\mathcal{L}^{1,ji}X}_{s-} + D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle\, e_j]\,\mathrm{d}s \\ &+ \int_0^t D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle\, e_j\,\mathrm{d}s + \int_0^t D_1\widehat{\mathcal{L}^{1,ji}X}_{s-}\,\mathrm{d}s + \widehat{\mathcal{M}}_t. \end{aligned} \quad (23)$$

The final form of the gain $G^{(\mathcal{L})}$ is obtained by identifying the terms with locally finite variations in the decompositions (22) and (23). The innovation form of the filter $\widehat{\mathcal{L}^{1,ji}X}_t$ in (13d) is deduced by replacing the gain $G^{(\mathcal{L})}$ in (21). □

2.4 Zakai filters

In this part of the paper, we derive the so-called Zakai filters associated with the statistics in (3). A technique of change of measure is used. We recall some basic facts which are borrowed from (Bremaud 1981, Chap VI) for instance. Assume that

$$\mu(i) := \langle \mathbf{1}, D_1 e_i\rangle > 0, \quad i = 1, \ldots, n.$$

Then, we have

$$0 < \min_i \mu(i) \leq \lambda_t = \langle \mathbf{1}, D_1 X_{t-}\rangle \leq \sum_i \mu(i)$$

so that

$$\frac{1}{\sum_i \mu(i)} \leq \frac{1}{\lambda_t} \leq \frac{1}{\min_i \mu(i)}. \tag{24}$$

Let us consider the likelihood ratio

$$\overline{L}_t \;:=\; \exp\left(\int_0^t \ln\frac{1}{\lambda_s}\mathrm{d}N_s + \int_0^t (\lambda_s - 1)\,\mathrm{d}s\right) = \prod_{0<s\leq t} (\frac{1}{\lambda_s})^{\Delta N_s} \exp\left(\int_0^t (\lambda_s - 1)\,\mathrm{d}s\right)$$

which is a solution of the equation

$$\overline{L}_t = 1 + \int_0^t \overline{L}_{s-}(\frac{1}{\lambda_s} - 1)\,(\mathrm{d}N_s - \lambda_s \mathrm{d}s).$$

Then, $(\overline{L}_t)_{t\geq 0}$ is a $(\mathbb{P}, \mathbb{F})$-martingale from (24) and (R3). A new probability measure $\mathbb{P}_0$ on $(\Omega, \vee_{t\geq 0}\mathbb{F}_t)$ is defined by

$$\left.\frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{P}}\right|_{\mathbb{F}_t} := \overline{L}_t.$$

It results from Girsanov's theorem that, under $\mathbb{P}_0$, $(N_t)_{t\geq 0}$ is a $\mathbb{F}$-homogeneous Poisson process with intensity $\overline{\lambda}_s \equiv 1$. Therefore, the process $(n_t)_{t\geq 0}$ defined by

$$n_t := N_t - t \tag{25}$$

is a $(\mathbb{P}_0, \mathbb{F})$-martingale and must be thought of as the innovations martingale associated with $(N_t)_{t\geq 0}$ under the new probability measure $\mathbb{P}_0$.

Next, set

$$L_t := \frac{1}{\overline{L}_t}. \tag{26}$$

$(L_t)_{t\geq 0}$ is a solution of the equation

$$L_t = 1 + \int_0^t L_{s-}(\lambda_s - 1)\,\mathrm{d}n_s \tag{27}$$

and is a $(\mathbb{P}_0, \mathbb{F})$-martingale. We know that $\mathbb{P} \ll \mathbb{P}_0$ and

$$\left.\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_0}\right|_{\mathbb{F}_t} = L_t.$$

Under the mild condition stated at the beginning of this subsection, $\mathbb{P}$ and $\mathbb{P}_0$ are equivalent probability measures. The expectation with respect to $\mathbb{P}_0$ is denoted by $\mathbb{E}_0[\cdot]$. Any $\mathbb{F}^N$-conditional expectation, under $\mathbb{P}$, of an integrable $\mathbb{F}$-measurable random variable, may be computed from Bayes formula (e.g. see Elliott et al. (1995))

$$\mathbb{E}[Z \mid \mathbb{F}_t^N] = \frac{\mathbb{E}_0[ZL_t \mid \mathbb{F}_t^N]}{\mathbb{E}_0[L_t \mid \mathbb{F}_t^N]} \quad \mathbb{P} \text{ a.s.} \tag{28}$$

The above equation allows us to derive the Zakai filters $\mathbb{E}_0[ZL_t \mid \mathbb{F}_t^N]$ by the formula

$$\mathbb{E}_0[ZL_t \mid \mathbb{F}_t^N] = \mathbb{E}[Z \mid \mathbb{F}_t^N]\ \mathbb{E}_0[L_t \mid \mathbb{F}_t^N]. \tag{29}$$

The following final fact is used to derive the Zakai filters associated with the statistics defined in (3). It is known from (Bremaud 1981, R 8, p 174) that $\widehat{L}_t^0 := \mathbb{E}_0[L_t \mid \mathbb{F}_t^N]$ satisfies the equation

$$\widehat{L}_t^0 = 1 + \int_0^t \widehat{L}_{s-}^0(\widehat{\lambda}_s - 1)\,\mathrm{d}n_s \tag{30}$$

where $\widehat{\lambda}_t$ is the $\mathbb{F}^N$-stochastic intensity of $(N_t)_{t\geq 0}$ given in (11).

Theorem 6 *The processes $(n_t)_{t\geq 0}$ and $(L_t)_{t\geq 0}$ are defined in (25) and (26) respectively. For any $\mathbb{F}$-adapted integrable process $(Z_t)_{t\geq 0}$, $\sigma(Z_t)$ denotes the conditional expectation $\mathbb{E}_0[Z_tL_t \mid \mathbb{F}_t^N]$.*

1. ***Zakai filter for the state.*** *We have for any $t \geq 0$*

$$\sigma(X_t) = \widehat{X}_0 + \int_0^t Q\sigma(X_{s-})\,\mathrm{d}s + \int_0^t (D_1 - I)\sigma(X_{s-})\,\mathrm{d}n_s. \tag{31a}$$

2. ***Zakai filter for the sojourn time in e_i.*** *We have for any $t \geq 0$*

$$\sigma(\mathcal{O}_t^{(i)}X_t) = \int_0^t \left[Q\sigma(\mathcal{O}_{s-}^{(i)}X_{s-}) + \langle \sigma(X_{s-}), e_i\rangle\, e_i\right]\mathrm{d}s + \int_0^t (D_1 - I)\sigma(\mathcal{O}_{s-}^{(i)}X_{s-})\,\mathrm{d}n_s. \tag{31b}$$

3. ***Zakai filter for the numbers of specific jumps for the MAP.*** *We have for any $t \geq 0$*

$$\sigma(\mathcal{L}_t^{0,ji}X_t) = \int_0^t \left[Q\sigma(\mathcal{L}_{s-}^{0,ji}X_{s-}) + D_0(j,i)\langle \sigma(X_{s-}), e_i\rangle\, e_j\right]\mathrm{d}s + \int_0^t (D_1 - I)\sigma(\mathcal{L}_{s-}^{0,ji}X_{s-})\,\mathrm{d}n_s \tag{31c}$$

$$\begin{aligned}\sigma(\mathcal{L}_t^{1,ji}X_t) = &\int_0^t \left[Q\sigma(\mathcal{L}_{s-}^{1,ji}X_{s-}) + D_1(j,i)\langle \sigma(X_{s-}), e_i\rangle\, e_j\right]\mathrm{d}s \\ &+ \int_0^t \left[(D_1 - I)\sigma(\mathcal{L}_{s-}^{1,ji}X_{s-}) + D_1(j,i)\langle \sigma(X_{s-}), e_i\rangle\, e_j\right]\mathrm{d}n_s.\end{aligned} \tag{31d}$$

Proof. Let $G_s^{(i)}$ be the innovations gain defined in (18). We find from the product rule (R1) and using (30),(13b) and (R3), that

$$\begin{aligned}\widehat{L}_t^0\widehat{\mathcal{O}^{(i)}X}_t &= \int_0^t \widehat{L}_{s-}^0\,\mathrm{d}\widehat{\mathcal{O}^{(i)}X}_s + \int_0^t \widehat{\mathcal{O}^{(i)}X}_{s-}\,\mathrm{d}\widehat{L}_s^0 + \sum_{0<s\leq t}\Delta\widehat{L}_s^0\Delta\widehat{\mathcal{O}^{(i)}X}_s \\ &= \int_0^t \widehat{L}_{s-}^0\left[Q\widehat{\mathcal{O}^{(i)}X}_{s-} + \langle \widehat{X}_{s-}, e_i\rangle\right]\mathrm{d}s + \int_0^t \widehat{L}_{s-}^0 G_s^{(i)}(\mathrm{d}N_s - \widehat{\lambda}_s\,\mathrm{d}s) \\ &\quad + \int_0^t \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{L}_{s-}^0(\widehat{\lambda}_s - 1)\,\mathrm{d}n_s + \int_0^t G_s^{(i)}\widehat{L}_{s-}^0(\widehat{\lambda}_s - 1)\,\mathrm{d}N_s \\ &= \int_0^t \widehat{L}_{s-}^0\left[Q\widehat{\mathcal{O}^{(i)}X}_{s-} + \langle \widehat{X}_{s-}, e_i\rangle\right]\mathrm{d}s + \int_0^t \widehat{L}_{s-}^0 G_s^{(i)}\widehat{\lambda}_s\mathrm{d}n_s + \int_0^t \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{L}_{s-}^0(\widehat{\lambda}_s - 1)\mathrm{d}n_s.\end{aligned}$$

Since we know from (18) that $G_s^{(i)}\widehat{\lambda}_s = D_1\widehat{\mathcal{O}^{(i)}X}_{s-} - \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{\lambda}_s$, we obtain after some simplifications that

$$\widehat{L}_t^0\widehat{\mathcal{O}^{(i)}X}_t \;=\; \int_0^t \left[Q\widehat{L}_{s-}^0\widehat{\mathcal{O}^{(i)}X}_{s-} + \langle \widehat{L}_{s-}^0\widehat{X}_{s-}, e_i\rangle\right] \mathrm{d}s + \int_0^t (D_1 - I)\widehat{L}_{s-}^0\widehat{\mathcal{O}^{(i)}X}_{s-}\,\mathrm{d}n_s.$$

Using (29) and the notation introduced in Theorem 6, the equation above has the final form (31b). The other formulas are obtained in a quite similar way. □

SDEs in Theorem 6 are standard linear ordinary differential equations (ODE) between two jumps of $(N_t)_{t\geq 0}$. Therefore, a basic way to deal with the equations (31a-31d) is to integrate the linear ODEs over the interval of time between two jumps and to update the solution at the endpoint of the interval. For instance, the state filter $\sigma(X_t)$ is solution of

$$\frac{\mathrm{d}}{\mathrm{d}t}\,q_t \;=\; (Q - D_1 + I)q_t = (D_0 + I)q_t, \quad q_{t_{l-1}} := \sigma(X_{t_{l-1}})$$

in the interval $[t_{l-1}, t_l[$, where $(t_l)_{l\in\mathbb{N}}$ is the sequence of times of jump of $(N_t)_{t\geq 0}$ ($t_0 := 0$). Then, the solution at time of jump t_l is updated as follows

$$\Delta\sigma(X_{t_l}) = (D_1 - I)\sigma(X_{t_l-}) \Longrightarrow \sigma(X_{t_l}) = D_1\sigma(X_{t_l-}).$$

In this special case, it is easily seen that (31a) has an explicit solution given for $t > 0$ by

$$\sigma(X_t) = \exp(t)\exp\big(D_0(t - t_{N_t})\big)D_1\exp\big(D_0(t_{N_t} - t_{N_t-1})\big)\cdots D_1\exp\big(D_0 t_1\big)\widehat{X}_0$$

and the vector of conditional probabilities $\widehat{X}_t$ is for $t > 0$

$$\widehat{X}_t = \frac{\exp\big(D_0(t - t_{N_t})\big)D_1\exp\big(D_0(t_{N_t} - t_{N_t-1})\big)\cdots D_1\exp\big(D_0 t_1\big)\widehat{X}_0}{\mathbf{1}^\top\exp\big(D_0(t - t_{N_t})\big)D_1\exp\big(D_0(t_{N_t} - t_{N_t-1})\big)\cdots D_1\exp\big(D_0 t_1\big)\widehat{X}_0}.$$

Assume that K jumps of $(N_t)_{t\geq 0}$ have been observed at times $0 < t_1 < \cdots < t_K$ and set $t_0 := 0$. The solutions of (31a-31d) may be computed on the grid $\Pi := \{0, t_1, \ldots, t_K\}$ from the recursive formulas in Figure 2. We are now in position to discuss the comparison between the two procedures reported in Figures 1,2.

2.5 Comments on the comparison of the forward-backward and filter-based strategies

The two procedures involve the computation of the matrix exponential functions f_0, f_1 as well as their integrals. The uniformization method is known to be efficient for computing transient measures of continuous-time Markov processes involving matrix exponentials. This method is based on the following decomposition of the matrix exponential of a sub-generator D_0 (i.e. $i \neq j$ $D_0(i,j) \geq 0$ and $(\mathbf{1}^\top D_0)(i) \leq 0$ for any i).

$$\exp(D_0 t) = \sum_{k=0}^{+\infty}\exp(-ut)\frac{u^k}{k!}P^k \quad \text{with } P := I + \frac{1}{u}D_0, \text{ and } u > \max_i(-D_0(i,i)).$$

$f_0(x) := \exp(D_0 x)$ and $f_1(x) := D_1 \exp(D_0 x)$; for $l = 1, \ldots, K$, $\Delta t_l := t_l - t_{l-1}$ with $t_0 := 0$.

for $l = 1, \ldots, K$

$$
\begin{aligned}
\sigma(X_{t_l}) &= f_1(\Delta t_l)\, \sigma(X_{t_{l-1}}) \\
\sigma(\mathcal{L}^{0,ji}_{t_l} X_{t_l}) &= f_1(\Delta t_l) \sigma(\mathcal{L}^{0,ji}_{t_{l-1}} X_{t_{l-1}}) + \int_{t_{l-1}}^{t_l} f_1(t_l - s) e_j\, D_0(j,i)\, e_i^\top f_0(s - t_{l-1})\, \mathrm{d}s\, \sigma(X_{t_{l-1}}) \\
\sigma(\mathcal{L}^{1,ji}_{t_l} X_{t_l}) &= f_1(\Delta t_l) \sigma(\mathcal{L}^{1,ji}_{t_{l-1}} X_{t_{l-1}}) + e_j\, D_1(j,i)\, e_i^\top f_0(\Delta t_l) \sigma(X_{t_{l-1}}) \\
\sigma(\mathcal{O}^{(i)}_{t_l} X_{t_l}) &= f_1(\Delta t_l)\, \sigma(\mathcal{O}^{(i)}_{t_{l-1}} X_{t_{l-1}}) + \int_{t_{l-1}}^{t_l} f_1(t_l - s) e_i e_i^\top f_0(s - t_{l-1})\, \mathrm{d}s\, \sigma(X_{t_{l-1}})
\end{aligned}
$$

Comment. The factor $\exp(\Delta t_l)$ is omitted in the equations above, because the estimates at a fixed instant of D_0, D_1 from (10) only require the knowledge of the filters up to a constant.

Figure 2: Recursive implementation of the filters

The numerical interests are in the facts that (1) matrix P is a non-negative (primitive) matrix, so that only non-negative real numbers are involved in the computations, (2) a robust computation of the Poisson probabilities may be carried out and (3) the level of truncation of the series above may be a priori controlled. We refer to Stewart (1994) for complete details.

Now we turn to the differences between the two procedures. The weakness of the filter-based approach is its computational cost. Indeed, we have to deal with a specific recursive equation for each statistic involved in the re-estimation formulas (10). So that, the computational cost is $O(n^4)$ in the number of parameters to be estimated. The corresponding cost for the forward-backward strategy is only $O(n^2)$. However, when fitting data to an MAP model, the experiments clearly show that the number n of states to be considered for the hidden Markov process $(X_t)_{t \geq 0}$ should be relatively small (n is not found to be greater than 3 in Klemm et al. (2003)). Then, it can be expected in some practical situations than the difference of computational complexity will be not too large.

It is clear from Figure 1 that the sequence of vectors $\alpha_l, l = 0, \ldots, K$ provided by the forward recursion have to be stored before performing the backward recursion. Therefore, the storage cost is $O(K)$ in the number of observed data. Now, the storage cost for the recursive procedure in Figure 2 does not depend on the number K of observations. We only have to store the previous estimate of each statistics. From this point of view, the filter-based approach must be preferred to the forward-backward one when processing large data-sets. This is the case in the experiments reported in Klemm et al. (2003), where trace files of $200,000$ arrivals are considered. Thus, a recursive implementation of the EM-algorithm has the advantage that the number of observations is not fixed so that on-line estimation may be considered, and this number may be as large as possible. Finally, the various filters in the filter-based scheme are decoupled and are suitable for parallel implementation on a multiprocessor system.

Large experiments on the use of the EM-algorithm have been carried out in Lang and Arthur (1996), Asmussen et al. (1996), Rydén (1996), Klemm et al. (2003) with the forward-backward strategy. Since the differences between the two procedures are mainly on the order of complexity, here it is not intended to provide further numerical experiments. The appealing properties and main drawbacks of the EM-algorithm in estimation problem are well-known. The implementation is relatively easy, the algorithm is robust but its convergence – if convergence takes place – is slow. We refer to Wu (1983) for a discussion on the convergence rate of EM.

3 Conclusion

When only the counting process is observed in a Markovian Arrival Process, parameter estimation may be carried out with the help of the EM-algorithm. Then, filters for various statistics associated with this model must be computed. With a view to implementing the filter-based strategy popularized by Elliott, the filters and an unnormalized/Zakai form of the filters are shown to be the solution of stochastic differential equations. As a result, we obtain recursive computational procedures for filtering. The case of batch arrivals, that is of BMAPs, may easily be included in the discussion. We just have to consider the observation process $(N_t)_{t\geq 0}$ as a multivariate point process of arrivals.

References

Asmussen, S. (1996). Phase-type distributions and related point processes: Fitting and recent advances. In Chakravarty, S. and Alfa, A., editors, *Matrix Analytic methods in Stochastic Models*, pages 137–149. Marcel Dekker.

Asmussen, S. (2000). Matrix-analytic models and their analysis. *Scand. J. Statist.*, 27:193–226.

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

Boucheron, S. and Gassiat, E. (2005). An information theoretic perspective on order estimation. In O. Cappé, E. Moulines, and T. Rydén, editors, *Inference in Hidden Markov Models*. Springer.

Bremaud, P. (1981). *Point Processes and Queues*. Springer Series in Statistics. Springer.

Breuer, L. (2002). An EM algorithm for batch arrival processes and its comparison to a simpler estimation procedure. *Ann. Oper. Res.*, 112:123–138.

Elliott, R. J., Aggoun, L., and Moore, J. (1995). *Hidden Markov Models*. Springer.

Ephraim, Y. and Merhav, N. (2002). Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48:1518–1569.

Fischer, W. and Meier-Hellstern, K. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171.

Goseva-Popstojanova, K. and Trivedi, K. S. (2001). Architecture-based approach to reliability assessment of software systems. *Performance Evaluation*, 45:179–204.

Gravereaux, J.-B. and Ledoux, J. (2004). Poisson approximation for some point processes in reliability. *Adv. in Appl. Probab.*, 36:455–470.

Klebaner, F. C. (1998). *Introduction to Stochastic Calculus with Applications*. Imperial College Press.

Klemm, A., Lindemann, C., and Lohmann, M. (2003). Modeling IP traffic using the Batch Markovian Arrival Process. *Performance Evaluation*, 54:149–173.

Kloeden, P. E., Platen, E., and Schurz, H. (1994). *Numerical Solution of SDE through Computer Experiments*. Springer-Verlag.

Lang, A. and Arthur, J. (1996). Parameter estimation for phase-type distributions. In Chakravarty, S. and Alfa, A., editors, *Matrix Analytic methods in Stochastic Models*, pages 151–206. Marcel Dekker.

Littlewood, B. (1975). A reliability model for systems with Markov structure. *Appl. Statist.*, 24:172–177.

Neuts, M. F. (1989). *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Dekker Inc., New-York and Basel.

Neuts, M. F. (1995). Matrix-analytic methods in queueing theory. In J. H. Dshalalow, editor, *Advances in Queueing*, Probability and Stochastic Series, pages 265–292. CRC Press.

Rydén, T. (1996). An EM-algorithm for estimation in Markov-modulated Poisson process. *Comput. Statist. Data Anal.*, 21:431–447.

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov chains*. Princeton University.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103.