Recursive filters for partially observable finite Markov chains

James Ledoux

Recursive filters for partially observable finite Markov chains

James Ledoux*

24 august 2005

Abstract

In this note, we consider discrete-time finite Markov Chains and assume that they are only partly observed. We obtain finite-dimensional normalized filters for basic statistics associated with such processes. Recursive equations for these filters are derived by means of simple computations involving conditional expectations. An application to the estimation of parameters for the so-called discrete-time Batch Markovian Arrival Processes is outlined.

KEYWORDS: time-inhomogeneous Markov chain; hidden Markov chain; D-BMAP; EM-algorithm
AMS 60J10,93E11

1 Introduction

Let $(X, Y) := (X_t, Y_t)_{t\geq 0}$ be a bivariate discrete-time finite (inhomogeneous) Markov chain. Only the second component Y is supposed observable. Y is often called the observation process and X the "hidden" process. The special case of hidden Markov models is well known. A Hidden Markov Chain (HMC) is a bivariate Markov chain for which the hidden component is also a Markov chain. This class of Markovian models has proved to be useful in many ares of probabilistic modeling, including speech process (e.g. see [15] and the references therein).

At time t, the available data on (X, Y) consists of the sequence of values $Y_0, \ldots, Y_t$. A standard problem is to find from the data the best estimate, at time t, of quantities related to the unobserved process X. Such a problem is often referred to as a filtering problem. In this paper, finite-dimensional filters for basic statistics related to X and Y are obtained in a recursive form. Our only assumption is that (X, Y) is a finite Markov chain. The derivation of our main result is based on basic computations on conditional expectations.

*IRMAR UMR 6625 & INSA de Rennes, 20 avenue des Buttes de Coësmes, CS14315, 35043 Rennes cedex, France

Specializing our result to time-homogeneous dependent hidden Markov chains, we find a renormalized form of the Zakaï filters obtained by Elliott in [7, 8]. The change of measure technique used by Elliott has the advantage of providing an unified framework for the computation of filters for hidden Markov chains. In particular, it works for hidden Markov chains continuous in both time and space [8]. Note that to be suitable, such a change of measure must preserve the Markov property of the hidden component X. In the context of discrete time and space, we believe our approach is more direct and elementary.

In this paper, the number of 'states' of X 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. We do not address this fundamental issue here. In the context of hidden Markov chains, the so-called 'order estimation problem' is surveyed in [10, Section VIII]. Recent progress in this direction is reported in [11] and [5]. We refer the reader to these papers and the references therein for details.

The paper is organized as follows. In section 2, we state our main result. Then, in Section 3, we specialize the result to the time-inhomogeneous dependent Hidden Markov chains. We also mention the connection of our result with the so-called innovation approach for filtering. In Section 4, we briefly describe parameter estimation for the discrete-time counterpart of the Batch Markovian Arrival Process. Here, a filter-based form of the EM-algorithm is used. Section 5 is devoted to the proof of our main result. Finally, we present our conclusions in Section 6.

1.1 Main notation

Throughout, we will use the following notation.

- By convention, vectors are column vectors. Row vectors are denoted by means of the transpose operator $(.)^{\mathsf{T}}$.
- $\mathbf{1}$ and $\mathbf{0}$ are vectors in which each entry is equal to one or 0, respectively. Their dimensions are defined by the context.
- $\boldsymbol{e}_i$ and $\boldsymbol{f}_j$ are the ith vector of the canonical basis of $\mathbb{R}^n$ and the jth vector of the canonical basis of $\mathbb{R}^m$, respectively.
- For any vector (or scalar) $\boldsymbol{a}$ and scalar $b \in \mathbb{R}$,

$$(\boldsymbol{a}/b)^+ := \begin{cases} \boldsymbol{a}/b & b \neq 0 \\ \mathbf{0}, & b = 0. \end{cases}$$

With this notation, $(b/b)^+$ is the indicator function of the set $\{b \neq 0\}$.

2 Recursive filters for partially observable finite Markov chains

In this section, (X, Y) is only assumed to be a finite Markov chain. The state space of (X, Y) is assumed to be the finite set $\mathscr{E} \times \mathscr{F}$ where $\mathscr{E} := \{\boldsymbol{e}_i, i = 1, \ldots, n\}$ and $\mathscr{F} := \{\boldsymbol{f}_k, k = 1, \ldots, m\}$. Note that with our conventions,

$$1_{\{X_t = \boldsymbol{e}_i\}} = \langle X_t, \boldsymbol{e}_i \rangle = \boldsymbol{e}_i^{\mathsf{T}} X_t, \quad X_t = \sum_{i=1}^{n} \langle X_t, \boldsymbol{e}_i \rangle \boldsymbol{e}_i \quad \mathbf{1}^{\mathsf{T}} X_t = 1$$

where $1_{\{\cdot\}}$ is the indicator function of the set $\{\cdot\}$ and $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean scalar product. The corresponding relations also hold for the observed process Y. All processes are supposed to be defined on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We denote by $\mathbb{F}^X = (\mathbb{F}_t^X)_{t \geq 0}$, $\mathbb{F}^Y = (\mathbb{F}_t^Y)_{t \geq 0}$, $\mathbb{F}^{X,Y} = (\mathbb{F}_t^{X,Y})_{t \geq 0}$ the internal filtrations of processes X, Y and (X, Y) respectively. That is,

$$\mathbb{F}_t^X := \sigma(X_s,\, s \leq t), \quad \mathbb{F}_t^Y := \sigma(Y_s,\, s \leq t), \quad \mathbb{F}_t^{X,Y} := \sigma(X_s, Y_s,\, s \leq t).$$

These filtrations are assumed to be complete, that is every σ-algebra contains all the sets of $\mathcal{F}$ of P-probability zero.

In this paper, we deal with the following statistics of the partially observed Markov chain, for any time $t \geq 1$ and $i, j = 1, \ldots, n$:

- the number of jumps of X from $\boldsymbol{e}_i$ to $\boldsymbol{e}_j$ up to time t, *i.e.*

$$N_t^{ji} := \sum_{l=1}^{t} \langle X_l, \boldsymbol{e}_j \rangle \langle X_{l-1}, \boldsymbol{e}_i \rangle; \tag{2.1}$$

- the number of visits of X to state $\boldsymbol{e}_i$ up to time $t-1$, *i.e.*

$$\mathcal{O}_t^{(i)} := \sum_{l=1}^{t} \langle X_{l-1}, \boldsymbol{e}_i \rangle; \tag{2.2}$$

- the number of times, up to time t, (X, Y) jumps to state $(\boldsymbol{f}_k, \boldsymbol{e}_j)$ given that X is in state $\boldsymbol{e}_i$, at the previous time, *i.e.*

$$\mathscr{L}_t^{k,ji} := \sum_{l=1}^{t} \langle Y_l, \boldsymbol{f}_k \rangle \langle X_l, \boldsymbol{e}_j \rangle \langle X_{l-1}, \boldsymbol{e}_i \rangle, \quad k = 1, \ldots, m. \tag{2.3}$$

For these scalar random variables, we are interested in their quadratic mean estimates from the data $(Y_0, Y_1, \ldots, Y_t)$:

$$\widehat{N^{ji}}_t := \mathrm{E}[N_t^{ji} \mid \mathbb{F}_t^Y], \quad \widehat{\mathcal{O}^{(i)}}_t := \mathrm{E}[\mathcal{O}_t^{(i)} \mid \mathbb{F}_t^Y] \quad \text{and} \quad \widehat{\mathscr{L}^{k,ji}}_t := \mathrm{E}\big[\mathscr{L}_t^{k,ji} \mid \mathbb{F}_t^Y\big]. \tag{2.4}$$

In fact, we find that $\widehat{X}_t := \mathrm{E}[X_t \mid \mathbb{F}_t^Y]$ and the filters defined by

$$\begin{aligned}\widehat{N^{ji}X}_t &:= \mathrm{E}\big[N_t^{ji}X_t \mid \mathbb{F}_t^Y\big],\\ \widehat{\mathcal{O}^{(i)}X}_t &:= \mathrm{E}\big[\mathcal{O}_t^{(i)}X_t \mid \mathbb{F}_t^Y\big]\\ \widehat{\mathscr{L}^{k,ji}X}_t &:= \mathrm{E}\big[\mathscr{L}_t^{k,ji}X_t \mid \mathbb{F}_t^Y\big],\end{aligned}$$

for any $t \geq 0$, each satisfies a recursive equation. Since $\mathbf{1}^{\mathsf{T}}X_t = 1$, the filters in (2.4) can be written

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

For any $t \geq 0$, we define the $n \times m$ random matrices $\boldsymbol{S}(t; X_t, Y_t)$ as follows: for any $j = 1, \ldots, n$ and $k = 1, \ldots, m$, its (j,k)th entry is

$$\begin{aligned}\boldsymbol{S}(t; X_t, Y_t)(j,k) &:= \mathrm{E}\big[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid X_t, Y_t\big]\\ &= \mathrm{P}\big((X_{t+1}, Y_{t+1}) = (\boldsymbol{e}_j, \boldsymbol{f}_k) \mid X_t, Y_t\big).\end{aligned} \tag{2.5}$$

Then, the Markov property of (X, Y), the equality

$$X_{t+1}{Y_{t+1}}^{\mathsf{T}} = \sum_{j=1}^{n}\sum_{k=1}^{m}\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \boldsymbol{e}_j\boldsymbol{f}_k^{\mathsf{T}}$$

and (2.5) allow us to write

$$\begin{aligned}\mathrm{E}\big[X_{t+1}{Y_{t+1}}^{\mathsf{T}} \mid \mathbb{F}_t^{X,Y}\big] &= \mathrm{E}\left[X_{t+1}{Y_{t+1}}^{\mathsf{T}} \mid X_t, Y_t\right]\\ &= \sum_{j=1}^{n}\sum_{k=1}^{m}\mathrm{E}\left[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid X_t, Y_t\right]\boldsymbol{e}_j\boldsymbol{f}_k^{\mathsf{T}}\\ &= \boldsymbol{S}(t; X_t, Y_t), \quad t \geq 0.\end{aligned} \tag{2.6}$$

Since $\mathbf{1}^{\mathsf{T}}X_t = 1$, we deduce from (2.6) that, for every $t \geq 0$,

$$\mathrm{E}\big[Y_{t+1} \mid \mathbb{F}_t^{X,Y}\big] = \mathrm{E}\left[Y_{t+1} \mid X_t, Y_t\right] = \boldsymbol{S}(t; X_t, Y_t)^{\mathsf{T}}\mathbf{1}. \tag{2.7}$$

Let $(U_t)_{t\geq 0}$ be any sequence of integrable random variables. For any t, set $\widehat{U}_t := \mathrm{E}[U_t \mid \mathbb{F}_t^Y]$. In an abuse of notation, we denote by $\boldsymbol{S}(t;\widehat{U}_t,Y_t)$ the $n\times m$ random matrix

$$\boldsymbol{S}(t;\widehat{U}_t,Y_t) := \sum_{i=1}^{n} \boldsymbol{S}(t;\boldsymbol{e}_i,Y_t)\ \langle \widehat{U}_t,\boldsymbol{e}_i\rangle \tag{2.8}$$

Since $\boldsymbol{S}(t;X_t,Y_t)=\sum_{i=1}^n \boldsymbol{S}(t;\boldsymbol{e}_i,Y_t)\langle X_t,\boldsymbol{e}_i\rangle$ and $\mathrm{E}\big[\langle X_t,\boldsymbol{e}_i\rangle \mid \mathbb{F}_t^Y\big] = \langle \mathrm{E}[X_t\mid \mathbb{F}_t^Y],\boldsymbol{e}_i\rangle$, the random matrix $\boldsymbol{S}(t;\widehat{X}_t,Y_t)$ is a conditional expectation:

$$\boldsymbol{S}(t;\widehat{X}_t,Y_t) = \mathrm{E}\left[\boldsymbol{S}(t;X_t,Y_t)\mid \mathbb{F}_t^Y\right]. \tag{2.9}$$

Theorem 2.1 *1.* (Estimator for the state.) *For $k=1,\dots,m$, let $\boldsymbol{p}_k$ be the n-dimensional vector defined by: $\boldsymbol{p}_k(i) := \mathrm{P}(X_0=\boldsymbol{e}_i, Y_0=\boldsymbol{f}_k), i=1,\dots,n$. For any $t\geq 0$, we have*

$$\widehat{X}_0 = \sum_{k=1}^{m}\left(\frac{\boldsymbol{p}_k}{\mathbf{1}^\mathsf{T}\boldsymbol{p}_k}\right)^+\langle Y_0,\boldsymbol{f}_k\rangle,\quad \widehat{X}_{t+1} = \sum_{k=1}^{m}\left(\frac{\boldsymbol{S}(t;\widehat{X}_t,Y_t)\boldsymbol{f}_k}{\mathbf{1}^\mathsf{T}\boldsymbol{S}(t;\widehat{X}_t,Y_t)\boldsymbol{f}_k}\right)^+\langle Y_{t+1},\boldsymbol{f}_k\rangle. \tag{2.10}$$

2. (Estimator for the number of visits to $\boldsymbol{e}_i$.) *Set $\mathcal{O}_0^{(i)} := 0$ and, for $t\geq 1$, define $\mathcal{O}_t^{(i)}$ according to (2.2). We have $\widehat{\mathcal{O}^{(i)}X}_0 = 0$ and, for any $t\geq 0$,*

$$\widehat{\mathcal{O}^{(i)}X}_{t+1} = \sum_{k=1}^{m}\left(\frac{\boldsymbol{S}(t;\widehat{\mathcal{O}^{(i)}X}_t,Y_t)\boldsymbol{f}_k + \boldsymbol{S}(t;\boldsymbol{e}_i,Y_t)\boldsymbol{f}_k\langle\widehat{X}_t,\boldsymbol{e}_i\rangle}{\mathbf{1}^\mathsf{T}\boldsymbol{S}(t;\widehat{X}_t,Y_t)\boldsymbol{f}_k}\right)^+\langle Y_{t+1},\boldsymbol{f}_k\rangle. \tag{2.11}$$

3. (Estimator for the number of jumps of X from $\boldsymbol{e}_i$ to $\boldsymbol{e}_j$.) *Set $N_0^{ji}=0$ and, for $t\geq 1$, define N_t^{ji} according to (2.1). We have $\widehat{N^{ji}X}_0=0$ and, for any $t\geq 0$,*

$$\widehat{N^{ji}X}_{t+1} = \sum_{k=1}^{m}\left(\frac{\boldsymbol{S}(t;\widehat{N^{ji}X}_t,Y_t)\boldsymbol{f}_k + \boldsymbol{e}_j^\mathsf{T}\boldsymbol{S}(t;\boldsymbol{e}_i,Y_t)\boldsymbol{f}_k\langle\widehat{X}_t,\boldsymbol{e}_i\rangle\boldsymbol{e}_j}{\mathbf{1}^\mathsf{T}\boldsymbol{S}(t;\widehat{X}_t,Y_t)\boldsymbol{f}_k}\right)^+\langle Y_{t+1},\boldsymbol{f}_k\rangle. \tag{2.12}$$

4. (Estimator for the number of joint transitions.) *Set $\mathscr{L}_0^{k,ji} := 0$ and for $t\geq 1$, define $\mathscr{L}_t^{k,ji}$ according to (2.3). We have: $\widehat{\mathscr{L}^{k,ji}X}_0 = 0$ and, for any $t\geq 0$,*

$$\begin{aligned}\widehat{\mathscr{L}^{k,ji}X}_{t+1} = \sum_{l=1}^{m}&\left(\frac{\boldsymbol{S}(t;\widehat{\mathscr{L}^{k,ji}X}_t,Y_t)\boldsymbol{f}_l}{\mathbf{1}^\mathsf{T}\boldsymbol{S}(t;\widehat{X}_t,Y_t)\boldsymbol{f}_l}\right)^+\langle Y_{t+1},\boldsymbol{f}_l\rangle\\ &+\left(\frac{\boldsymbol{e}_j^\mathsf{T}\boldsymbol{S}(t;\boldsymbol{e}_i,Y_t)\boldsymbol{f}_k}{\mathbf{1}^\mathsf{T}\boldsymbol{S}(t;\widehat{X}_t,Y_t)\boldsymbol{f}_k}\langle\widehat{X}_t,\boldsymbol{e}_i\rangle\boldsymbol{e}_j\right)^+\langle Y_{t+1},\boldsymbol{f}_k\rangle.\end{aligned} \tag{2.13}$$

The equalities (2.10)–(2.13) are equalities up to a set of P-probability zero.

3 Hidden Markov chains

Let us consider a bivariate Markov chain $(X, Y) := (X_t, Y_t)_{t\geq 0}$. X_0 and Y_0 are assumed to be independent random variables. Suppose that the transition probabilities of (X, Y) satisfy

$$\begin{aligned}&\mathrm{P}\big((X_{t+1}, Y_{t+1}) = (\boldsymbol{e}_j, \boldsymbol{f}_k) \mid X_t, Y_t\big) = \mathrm{P}\big((X_{t+1}, Y_{t+1}) = (\boldsymbol{e}_j, \boldsymbol{f}_k) \mid X_t\big)\\ &\iff \mathrm{E}\big[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid X_t, Y_t\big] = \mathrm{E}\big[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid X_t\big]\\ &t \geq 0, \quad k = 1, \ldots, m \quad j = 1, \ldots, n.\end{aligned}$$

For any $t \geq 0$, let us define the matrices $\boldsymbol{D}_k(t)$, $k = 1, \ldots, m$ via their components $\boldsymbol{D}_k(t; \cdot, \cdot)$ by

$$\mathrm{P}\big((X_{t+1}, Y_{t+1}) = (\boldsymbol{e}_j, \boldsymbol{f}_k) \mid X_t = \boldsymbol{e}_i\big) = \boldsymbol{D}_k(t; j, i) \qquad i, j = 1, \ldots, n. \tag{3.1}$$

It is immediate that X is a Markov chain with transition matrices $(\boldsymbol{P}(t))_{t\geq 0}$ given by

$$\boldsymbol{P}(t) := \sum_{k=1}^{m} \boldsymbol{D}_k(t).$$

For such a Markov chain (X, Y), the random matrix $\boldsymbol{S}(t; X_t, Y_t)$ defined by (2.5) does not depend on Y_t and (2.6) becomes

$$\mathrm{E}\big[X_{t+1}{Y_{t+1}}^{\mathsf{T}} \mid \mathbb{F}_t^{X,Y}\big] = \mathrm{E}\big[X_{t+1}{Y_{t+1}}^{\mathsf{T}} \mid X_t\big] = \boldsymbol{S}(t; X_t), \quad t \geq 0, \tag{3.2}$$

where $\boldsymbol{S}(t; X_t)$ is the following $n \times m$ random matrix

$$\boldsymbol{S}(t; X_t) := \big(\boldsymbol{D}_1(t)X_t \ \cdots \ \boldsymbol{D}_m(t)X_t\big) \quad t \geq 0, \tag{3.3}$$

i.e. the jth column of $\boldsymbol{S}(t; X_t)$ is given by $\boldsymbol{D}_j(t)X_t$. Therefore, the conditional expectation in (2.7) can be written as

$$\mathrm{E}\big[Y_{t+1} \mid \mathbb{F}_t^{X,Y}\big] = \mathrm{E}\left[Y_{t+1} \mid X_t\right] = \boldsymbol{S}(t; X_t)^{\mathsf{T}}\mathbf{1} = \boldsymbol{G}(t)X_t \quad \text{for every } t \geq 0, \tag{3.4}$$

where $\boldsymbol{G}(t)$ is the $m \times n$ matrix

$$\boldsymbol{G}(t) := \begin{pmatrix} \mathbf{1}^{\mathsf{T}}\boldsymbol{D}_1(t) \\ \vdots \\ \mathbf{1}^{\mathsf{T}}\boldsymbol{D}_m(t) \end{pmatrix}.$$

Since $\mathrm{E}[X_{t+1} \mid X_t] = \boldsymbol{P}(t)X_t$, the Markov chain X has the representation

$$X_{t+1} = \boldsymbol{P}(t)X_t + V_{t+1}, \qquad t \geq 0, \tag{3.5}$$

where $V := (V_t)_{t\geq 1}$ is a $\mathbb{F}^X$-martingale difference (i.e. V_{t+1} is an integrable $\mathbb{F}^X_{t+1}$-measurable random variable satisfying $\mathrm{E}[V_{t+1} \mid \mathbb{F}^X_t] = \mathbf{0}$). Finally, from (3.5) and (3.4), the processes X and Y have the following representation for $t \geq 0$, where V is an $\mathbb{F}^X$-martingale difference and $W := (W_t)_{t\geq 1}$ is an $\mathbb{F}^{X,Y}$-martingale difference:

$$\left\{ \begin{array}{rcl} X_{t+1} & = & \boldsymbol{P}(t)X_t + V_{t+1} \\ Y_{t+1} & = & \boldsymbol{G}(t)X_t + W_{t+1}. \end{array} \right. \tag{3.6}$$

The two processes V and W are sometimes called the state noise and the observation noise, respectively. In the common set-up in hidden Markov modeling, the noise processes V and W are assumed to be uncorrelated given X, that is

$$\mathrm{E}\left[V_{t+1}{W_{t+1}}^{\mathsf{T}} \mid X_t\right] = \mathbf{0}.$$

Using (3.2) and the representations (3.6), this last condition may be rewritten as

$$\begin{aligned} &\mathrm{E}\left[X_{t+1}{Y_{t+1}}^{\mathsf{T}} \mid X_t\right] = \mathrm{E}\left[X_{t+1} \mid X_t\right] \mathrm{E}[Y_{t+1} \mid X_t]^{\mathsf{T}} \\ \Longleftrightarrow\ &\boldsymbol{S}(t; X_t) = \boldsymbol{P}(t)X_t(\boldsymbol{G}(t)X_t)^{\mathsf{T}} = \boldsymbol{P}(t)\mathrm{diag}(X_t)\boldsymbol{G}(t)^{\mathsf{T}}. \end{aligned}$$

We refer the reader to [9] for results on filtering in such a context. In this section, we deal with hidden Markov models that only satisfy Property (3.1). Such processes are referred to as dependent hidden Markov chains. Filtering for time-homogeneous dependent hidden Markov chains was considered in [8, Section 2.10] and [7]. The following result is obtained as a corollary of Theorem 2.1. As mentioned above, the matrix $\boldsymbol{S}(t; X_t, Y_t)$ in (2.7) does not depend on Y_t and is replaced by the matrix $\boldsymbol{S}(t; X_t)$ defined by (3.3). Then, the random matrix $\boldsymbol{S}(t; \widehat{X}_t)$ is $\sum_{i=1}^n \boldsymbol{S}(t; \boldsymbol{e}_i)\langle \widehat{X}_t, \boldsymbol{e}_i\rangle$ according to (2.8), and

$$\begin{aligned} \mathbf{1}^{\mathsf{T}}\boldsymbol{S}(t; \widehat{X}_t)\boldsymbol{f}_k &= \sum_{i=1}^n \mathbf{1}^{\mathsf{T}}\boldsymbol{S}(t; \boldsymbol{e}_i)\boldsymbol{f}_k \ \langle \widehat{X}_t, \boldsymbol{e}_i\rangle \\ &= \sum_{i=1}^n \mathbf{1}^{\mathsf{T}}\boldsymbol{D}_k(t)\boldsymbol{e}_i \ \langle \widehat{X}_t, \boldsymbol{e}_i\rangle \quad \text{from (3.3)} \\ &= \mathbf{1}^{\mathsf{T}}\boldsymbol{D}_k(t)\widehat{X}_t. \end{aligned}$$

Corollary 3.1 *1.* (Estimator for the state.) *Let $\boldsymbol{x}_0$ be a stochastic vector corresponding to the probability distribution of the random variable X_0. We have*

$$\widehat{X}_0 = \boldsymbol{x}_0, \qquad \widehat{X}_{t+1} = \sum_{k=1}^m \left(\frac{\boldsymbol{D}_k(t)\widehat{X}_t}{\mathbf{1}^{\mathsf{T}}\boldsymbol{D}_k(t)\widehat{X}_t}\right)^+ \langle Y_{t+1}, \boldsymbol{f}_k\rangle, \quad t \geq 0. \tag{3.7}$$

2. (Estimator for the number of visits to $\boldsymbol{e}_i$.) *We have* $\widehat{\mathcal{O}^{(i)}X}_0 = 0$ *and, for* $t \geq 0$,

$$\widehat{\mathcal{O}^{(i)}X}_{t+1} = \sum_{k=1}^{m} \left(\frac{\boldsymbol{D}_k(t)\widehat{\mathcal{O}^{(i)}X}_t + \langle \widehat{X}, \boldsymbol{e}_i \rangle \boldsymbol{D}_k(t)\boldsymbol{e}_i}{\mathbf{1}^\mathsf{T} \boldsymbol{D}_k(t)\widehat{X}_t} \right)^+ \langle Y_{t+1}, \boldsymbol{f}_k \rangle.$$

3. (Estimator for the number of jumps of X from $\boldsymbol{e}_i$ to $\boldsymbol{e}_j$.) *We have* $\widehat{N^{ji}X}_0 = 0$ *and, for* $t \geq 0$,

$$\widehat{N^{ji}X}_{t+1} = \sum_{k=1}^{m} \left(\frac{\boldsymbol{D}_k(t)\widehat{N^{ji}X}_t + \boldsymbol{D}_k(t;j,i)\langle \widehat{X}_t, \boldsymbol{e}_i \rangle \, \boldsymbol{e}_j}{\mathbf{1}^\mathsf{T} \boldsymbol{D}_k(t)\widehat{X}_t} \right)^+ \langle Y_{t+1}, \boldsymbol{f}_k \rangle.$$

4. (Estimator for the number of joint transitions.) *We have* $\widehat{\mathscr{L}^{k,ji}X}_0 = 0$ *and, for* $t \geq 0$,

$$\widehat{\mathscr{L}^{k,ji}X}_{t+1} = \sum_{l=1}^{m} \left(\frac{\boldsymbol{D}_l(t)\widehat{\mathscr{L}^{k,ji}X}_t}{\mathbf{1}^\mathsf{T} \boldsymbol{D}_l(t)\widehat{X}_t} \right)^+ \langle Y_{t+1}, \boldsymbol{f}_l \rangle + \left(\frac{\boldsymbol{D}_k(t;j,i)}{\mathbf{1}^\mathsf{T} \boldsymbol{D}_k(t)\widehat{X}_t} \langle \widehat{X}_t, \boldsymbol{e}_i \rangle \boldsymbol{e}_j \right)^+ \langle Y_{t+1}, \boldsymbol{f}_k \rangle.$$

Remark 3.1 *1. A similar estimator for the state was derived by Aström [3] using Bayes's formula.*

2. We have obtained a renormalized form of the recursive formulas provided in [7] and [8, Section 2.10] for time-homogeneous hidden Markov chains. It is such a form that is numerically implemented in practice. The simple setting of discrete-time processes with finite state spaces allows us to provide a direct approach. However, the change-of-measure approach used in [7] and [8] is general enough to deal with much more complex hidden Markov models than the ones considered here. Note that, for discrete- or continuous-time hidden Markov chains, the measure change must preserve the Markov property of X.

Remark 3.2 *The recursive (3.7) may be rewritten as*

$$\widehat{X}_0 = \boldsymbol{x}_0, \qquad \widehat{X}_{t+1} = \boldsymbol{P}(t)\widehat{X}_t + K_t\big(Y_{t+1} - \boldsymbol{G}(t)\widehat{X}_t\big) \quad t \geq 0 \tag{3.8}$$

where K_t *is the* $\mathbb{F}_t^Y$*-measurable matrix-valued random variable defined by*

$$K_t := \begin{pmatrix} \left(\frac{\boldsymbol{D}_1(t)\widehat{X}_t}{\mathbf{1}^\mathsf{T} \boldsymbol{D}_1(t)\widehat{X}_t} \right)^+ & \cdots & \left(\frac{\boldsymbol{D}_m(t)\widehat{X}_t}{\mathbf{1}^\mathsf{T} \boldsymbol{D}_m(t)\widehat{X}_t} \right)^+ \end{pmatrix}$$

Indeed, (3.7) has the form

$$\widehat{X}_{t+1} = K_t Y_{t+1},$$

whence

$$\widehat{X}_{t+1} \;=\; K_t\big(Y_{t+1} - \boldsymbol{G}(t)\widehat{X}_t\big) + K_t\boldsymbol{G}(t)\widehat{X}_t.$$

Since

$$K_t = \big(\, \boldsymbol{D}_1(t)\widehat{X}_t \cdots \boldsymbol{D}_m(t)\widehat{X}_t \,\big)\ \mathrm{diag}\Big(\big(1/\mathbf{1}^\top \boldsymbol{D}_k(t)\widehat{X}_t\big)^+\Big)$$

and (see (3.4))

$$\boldsymbol{G}(t)\widehat{X}_t = \begin{pmatrix} \mathbf{1}^\top \boldsymbol{D}_1(t)\widehat{X}_t \\ \vdots \\ \mathbf{1}^\top \boldsymbol{D}_m(t)\widehat{X}_t \end{pmatrix}$$

we find that

$$\begin{aligned} K_t\boldsymbol{G}(t)\widehat{X}_t &= \big(\, \boldsymbol{D}_1(t)\widehat{X}_t \cdots \boldsymbol{D}_m(t)\widehat{X}_t \,\big)\ \left(\left(\frac{\mathbf{1}^\top \boldsymbol{D}_k(t)\widehat{X}_t}{\mathbf{1}^\top \boldsymbol{D}_k(t)\widehat{X}_t}\right)^+\right)_{k=1}^{m} \\ &= \sum_{k=1}^{m} \boldsymbol{D}_k(t)\widehat{X}_t \left(\frac{\mathbf{1}^\top \boldsymbol{D}_k(t)\widehat{X}_t}{\mathbf{1}^\top \boldsymbol{D}_k(t)\widehat{X}_t}\right)^+. \end{aligned}$$

Note that if $\mathbf{1}^\top \boldsymbol{D}_k(t)\widehat{X}_t = 0$ *then* $\boldsymbol{D}_k(t)\widehat{X}_t = \mathbf{0}$*, meaning that*

$$K_t\boldsymbol{G}(t)\widehat{X}_t \;=\; \sum_{k=1}^{m} \boldsymbol{D}_k(t)\widehat{X}_t.$$

Since $\boldsymbol{P}(t) = \sum_{k=1}^{m} \boldsymbol{D}_k(t)$*, the proof of (3.8) is complete.*

Conditioning with respect to $\mathbb{F}_t^Y$ *in (3.4), we obtain* $\boldsymbol{G}(t)\widehat{X}_t = \mathrm{E}[Y_{t+1} \mid \mathbb{F}_t^Y]$*. Hence, the random process* $(K_t)_{t\geq 0}$ *is the gain with respect to the innovation martingale* $(Y_t - \mathrm{E}[Y_t \mid \mathbb{F}_{t-1}^Y])_{t\geq 1}$*. Equation (3.8) is the Doob decomposition of the process* $\widehat{X}$ *and can be thought of as the 'projection of the decomposition (3.5) for* X *onto the internal filtration* $\mathbb{F}^Y$*'. A related reference is [18]. For each filter in Corollary 3.1, an analogue of (3.8) may be derived in a similar way.*

4 Parameter estimation for D-BMAPs

We consider the discrete-time counterpart of the Batch Markovian Arrival Process (D-BMAP) defined by Neuts [14]. For the purpose of this section, only time-homogeneous Markov chains are considered. A standard presentation of a D-BMAP is as follows. Let us consider a Markov chain $(X, N) := (X_t, N_t)_{t\geq 0}$ over the state space $\mathscr{E} \times \mathbb{N}$, with transition probabilities satisfying

$$\begin{aligned} &\mathrm{P}\big((X_{t+1}, N_{t+1}) = (\boldsymbol{e}_j, l+k) \mid (X_t, N_t) = (\boldsymbol{e}_i, l)\big) \\ &= \mathrm{P}\big((X_{t+1}, N_{t+1} - N_t) = (\boldsymbol{e}_j, k) \mid X_t = \boldsymbol{e}_i\big) = \boldsymbol{D}_k(j,i). \end{aligned}$$

for all $l \geq 0$, $t \geq 0$, $i, j = 1, \ldots, n$, and $k = 0, \ldots, m$. The other transition probabilities are 0. $N_0 = 0$ and the probability distribution of X_0 is denoted by $\boldsymbol{x}_0$. The above matrices $\boldsymbol{D}_k$ $(k = 0, \ldots, m)$, agree with those of Lucantoni's formalism for defining the finite BMAP [2]. The nonnegative number $\boldsymbol{D}_0(j, i)$ represents the probability that X jumps from state $\boldsymbol{e}_i$ to $\boldsymbol{e}_j$ with no arrival event. For $k \geq 1$, the entry $\boldsymbol{D}_k(j, i)$ is the probability that X jumps from state $\boldsymbol{e}_i$ to state $\boldsymbol{e}_j$ with the occurrence of k arrivals. Then X is a Markov chain with transition probability matrix $\boldsymbol{P} = \sum_{k=0}^{m} \boldsymbol{D}_k$. D-BMAPs have been found useful in numerous situations (e.g. see [4] and [1]). We point out the central role played by $\boldsymbol{D}_k$s for analyzing the number N_t of arrivals, up to time t, generated by a D-BMAP. For instance, the distribution function of N_t may be numerically evaluated using the following system of difference equations, where $\boldsymbol{x}(l, t) := (\mathrm{P}(N_t \leq l \mid X_0 = \boldsymbol{e}_i))_{i=1}^{n}$:

$$\boldsymbol{x}(l, t) = \sum_{k=0}^{l \wedge m} \boldsymbol{D}_k \, \boldsymbol{x}(l - k, t - 1), \quad l \geq 0, t \geq 1 \qquad \boldsymbol{x}(l, 0) = \boldsymbol{1}, \quad l \geq 0.$$

Let $\boldsymbol{\theta}$ be the nonnegative parameter vector

$$\boldsymbol{\theta} := \{\boldsymbol{D}_k(j, i), \quad k = 0, \ldots, m \quad i, j = 1, \ldots, n\}$$

which satisfies: $\sum_{k=0}^{m} \sum_{j=1}^{n} \boldsymbol{D}_k(j, i) = 1$ for $i = 1, \ldots, n$. The vector $\boldsymbol{\theta}$ must be estimated from the data. We stress that a D-BMAP (X, N) can be thought of as a time-homogeneous hidden Markov chain (X, Y) as defined in Section 3, for which $\mathrm{P}(Y_0 = \boldsymbol{f}_0) = 1$ and

$$\mathrm{P}\big((X_{t+1}, Y_{t+1}) = (\boldsymbol{e}_j, \boldsymbol{f}_k) \mid X_t = \boldsymbol{e}_i\big) = \boldsymbol{D}_k(j, i) \quad k = 0, \ldots, m \quad i, j = 1, \ldots, n.$$

Thus, $\boldsymbol{f}_k$ stands for a k-arrival(s) event $(k = 0, \ldots, m)$. In what follows, a D-BMAP is identified with the associated HMC (X, Y) defined above. The EM-algorithm is known to be a standard procedure for the statistical estimation of discrete-time hidden Markov chains [15].

Methods of statistical estimation for the Markovian arrival process has been recently developed in the continuous-time context. Specifically, the EM-algorithm has been used by Rydén [16] for the Markov Modulated Poisson Process, by Asmussen and its co-authors for the Phase-Type distributions (see [2] and references therein), by Lindemann and its co-authors for general Markovian arrival processes [12], and by Breuer [6] for BMAPs. The numerical experiments reported in these studies show that EM-algorithm works well in general. We mention that the number of states of X is assumed to be known in these works. Procedures for estimating the number n of states of X are discussed in [17] (a brief discussion on the order estimation was also presented in [6, p 124]).

We briefly explain the EM-algorithm for D-BMAPs. For a fixed parameter vector $\boldsymbol{\theta}$, the underlying probability measure and the associated expectation are denoted by $\mathrm{P}_{\boldsymbol{\theta}}$ and $\mathrm{E}_{\boldsymbol{\theta}}$, respectively. The random variable X_0 or its probability distribution $\boldsymbol{x}_0$ is assumed to

be known. Under $\mathrm{P}_{\boldsymbol{\theta}}$, the likelihood function for the complete data (X, Y), up to time t, is

$$L_t(\boldsymbol{\theta}; X, Y) := \prod_{l=1}^{t} \prod_{k=0}^{m} \prod_{i,j=1}^{n} \boldsymbol{D}_k(j,i)^{\langle Y_l, \boldsymbol{f}_k \rangle \langle X_l, \boldsymbol{e}_j \rangle \langle X_{l-1}, \boldsymbol{e}_i \rangle} \prod_{i=1}^{n} \boldsymbol{x}_0(i)^{\langle X_0, \boldsymbol{e}_i \rangle},$$

where $\boldsymbol{x}_0(i)$ is the ith component of $\boldsymbol{x}_0$. The parameter $\boldsymbol{\theta}$ is estimated from the observations $(Y_1, \ldots, Y_t)$, using the following iterative procedure. The vector $\boldsymbol{\theta}_p := \{\boldsymbol{D}_k^{(p)}(j,i),\ i,j = 1, \ldots, n;\ k = 0, \ldots, m\}$ denotes the estimate of $\boldsymbol{\theta}$ at the end of iteration p.

1. Initialization : Choose $\boldsymbol{\theta}_0$

2. Expectation step. Set $\boldsymbol{\theta} := \boldsymbol{\theta}_p$. Consider the so-called pseudo-log-likelihood function $\boldsymbol{\theta}^* \mapsto Q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$ defined by

$$Q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) := \mathrm{E}_{\boldsymbol{\theta}}\big[\log L_t(\boldsymbol{\theta}^*; X, Y) \mid \mathbb{F}_t^Y\big] = \sum_{k=0}^{m} \sum_{i,j=1}^{n} \log D_k^*(j,i)\, \widehat{\mathscr{L}^{k,ji}}_t + K \tag{4.1}$$

   where $\boldsymbol{\theta}^* := \{D_k^*(j,i),\ i,j = 1, \ldots, n;\ k = 0, \ldots, m\}$, K does not depend on $\boldsymbol{\theta}^*$ and $\widehat{\mathscr{L}^{k,ji}}_t := \mathrm{E}_{\boldsymbol{\theta}}[\mathscr{L}_t^{k,ji} \mid \mathbb{F}_t^Y]$.

3. Maximization step. Determine the $\boldsymbol{\theta}_{p+1}$ maximizing the function (4.1) under the constraints

$$\sum_{i=1}^{n} \widehat{\mathcal{O}^{(i)}}_t \sum_{k=0}^{m} \sum_{j=1}^{n} D_k^*(j,i) = t,$$

   where $\widehat{\mathcal{O}^{(i)}}_t := \mathrm{E}_{\boldsymbol{\theta}}[\mathcal{O}_t^{(i)} \mid \mathbb{F}_t^Y]$. We obtain

$$\boldsymbol{D}_k^{(p+1)}(j,i) = \frac{\widehat{\mathscr{L}^{k,ji}}_t}{\widehat{\mathcal{O}^{(i)}}_t}, \quad i,j = 1, \ldots, n \tag{4.2}$$

4. Return in step 2 until a stopping criterion is satisfied.

It is clear that the best estimate of the transition probability $\boldsymbol{P}(j,i)$ at the end of iteration $p+1$ is given by

$$\boldsymbol{P}^{(p+1)}(j,i) = \frac{\widehat{N^{ji}}_t}{\widehat{\mathcal{O}^{(i)}}_t} \qquad i,j = 1, \ldots, n.$$

The EM-algorithm produces a sequence of parameter vectors $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots$ such that [10]:

- the sequence of values of the observed likelihood function $(l_t(\boldsymbol{\theta}_p; Y))_p$ is nondecreasing (with equality if and only if $\boldsymbol{\theta}_{p+1} = \boldsymbol{\theta}_p$ under an identifiability condition), and

- under appropriate conditions, the sequence $(l_t(\boldsymbol{\theta}_p; Y))_p$ converges to a local maximum of the observed likelihood function.

Note that the zero entries of $\boldsymbol{D}_k$s are preserved by the procedure above.

As a result of the procedure above, we must compute the estimates in (4.2). The standard method is to use the Baum-Welch implementation of the maximization step of the EM-algorithm (also referred to as the "forward-backward" technique). This has been done in the previously mentioned works for continuous-time BMAPs [2], [6], [12]. In the filter-based approach pioneered by Elliott [7], [8], the estimates (4.2) are directly computed from the recursive equations given in Corollary 3.1. Therefore, the filter-based EM-procedure is very easily implemented and retains the well-established statistical properties of the EM-algorithm. The basic difference with the standard Baum-Welch method, is that only one pass through the data set is needed in the filter-based method. This approach was shown to be of value for specific continuous-time Markovian models in a recent paper [13].

We mention that a simple procedure for estimating the matrices $\boldsymbol{D}_k$ associated with a continuous-time BMAP, was numerically compared to the EM procedure in [6]. Breuer's procedure [6] was shown to be much faster and to require less storage than the Baum-Welch implementation of the EM-algorithm. Breuer provided no details on the computational complexity and the statistical properties of his procedure. A discrete-time version of this simple procedure could be designed. We mention that the storage space required by the filter-based implementation of the EM-algorithm is independent of the number of observations (indeed, only the filtered variables at each instant need to be stored in order to compute the variables at the next instant). The filter-based EM-algorithm has a slightly higher computational complexity than its Baum-Welch form. Therefore, Breuer's procedure should perform faster than ours. However, due to the simplicity of the filter-based EM-algorithm for D-BMAPs, the speed increase should not be very great. We will not go into further details of Breuer's procedure.

5 Proof of Theorem 2.1

A formula for the conditional expectation given $\mathbb{F}_t^Y$ $(t \geq 0)$ is proposed in the lemma below. Equation (5.5) is our basic formula for deriving the recursive equations reported in Theorem 2.1. It is based on the natural dynamics of the observations:

$$\mathbb{F}_{t+1}^Y = \mathbb{F}_t^Y \vee \sigma(Y_{t+1}).$$

Here, $\sigma(Y_{t+1})$ is a σ-subalgebra of $\mathcal{F}$ generated by the finite partition $\{A_1, \ldots, A_m\}$ of Ω with

$$A_k := \{Y_{t+1} = \boldsymbol{f}_k\}.$$

The conditional probability given A_k is defined, for any $B \in \mathcal{F}$, by

$$\mathrm{P}_{A_k}(B) := \left(\frac{1}{\mathrm{P}(A_k)}\right)^+ \mathrm{E}[1_{B\cap A_k}].$$

Then, for any integrable random variable X, we have

$$\mathrm{E}_{A_k}[X] = \left(\frac{1}{\mathrm{P}(A_k)}\right)^+ \mathrm{E}[X 1_{A_k}] \tag{5.1}$$

$$\text{and } \mathrm{E}[X \mid \sigma(Y_{t+1})] = \sum_{k=1}^{m} \mathrm{E}_{A_k}[X]\, 1_{A_k}. \tag{5.2}$$

Lemma 5.1 *Consider the filtration $\mathbb{F}^Y = (\mathbb{F}_t^Y)_{t\geq 0}$ of $\mathcal{F}$ and an integrable random variable X. The following relations hold for every $t \geq 0$:*

$$\mathrm{E}[X \mid \mathbb{F}_t^Y \vee \sigma(Y_{t+1})] = \sum_{k=1}^{m} \mathrm{E}_{A_k}[X \mid \mathbb{F}_t^Y]\, 1_{A_k} \tag{5.3}$$

$$\mathrm{E}[1_{A_k}\, X \mid \mathbb{F}_t^Y] = \mathrm{E}[1_{A_k} \mid \mathbb{F}_t^Y]\, \mathrm{E}_{A_k}[X \mid \mathbb{F}_t^Y] \tag{5.4}$$

$$\mathrm{E}[X \mid \mathbb{F}_t^Y \vee \sigma(Y_{t+1})] = \sum_{k=1}^{m} \left(\frac{\mathrm{E}[1_{A_k}\, X \mid \mathbb{F}_t^Y]}{\mathrm{E}[1_{A_k} \mid \mathbb{F}_t^Y]}\right)^+ 1_{A_k} \tag{5.5}$$

Remark 5.1 *Lemma 5.1 still holds when the random variable Y_{t+1} is replaced by any finite-valued random variable. A general form of Formula (5.4) is referred to as a conditional Bayes's formula in [8].*

Proof[. Proof of Lemma 5.1] Let t be a positive integer. Any element of the σ-algebra $\mathbb{F}_t^Y \vee \sigma(Y_{t+1})$ is a finite union of sets of the form $B \cap A_k$ for some $B \in \mathbb{F}_t^Y$. The summation on the right-hand side member of (5.3) is $\mathbb{F}_t^Y \vee \sigma(Y_{t+1})$-measurable. Hence, to derive (5.3), we must show that, for $k = 1, \ldots, m$,

$$\mathrm{E}[1_{B\cap A_k}\, X] = \mathrm{E}\big[1_{B\cap A_k} \sum_{i=1}^{m} 1_{A_i}\, \mathrm{E}_{A_i}[X \mid \mathbb{F}_t^Y]\big], \quad B \in \mathbb{F}_t^Y.$$

We can write

$$\begin{aligned}
\mathrm{E}[1_{B\cap A_k}\, X] &= \mathrm{P}(A_k)\, \mathrm{E}_{A_k}[1_B\, X] \quad \text{from (5.1)} \\
&= \mathrm{P}(A_k)\, \mathrm{E}_{A_k}\left[\mathrm{E}_{A_k}[1_B\, X \mid \mathbb{F}_t^Y]\right] \\
&= \mathrm{P}(A_k)\, \mathrm{E}_{A_k}\left[1_B \mathrm{E}_{A_k}[X \mid \mathbb{F}_t^Y]\right] \quad \text{since } B \in \mathbb{F}_t^Y \\
&= \mathrm{E}\left[1_B\, \mathrm{E}_{A_k}[X \mid \mathbb{F}_t^Y]\, 1_{A_k}\right] \quad \text{applying (5.1) to the r.v. } 1_B\, \mathrm{E}_{A_k}[X \mid \mathbb{F}_t^Y] \\
&= \mathrm{E}\big[1_{B\cap A_k} \sum_{i=1}^{m} 1_{A_i}\, \mathrm{E}_{A_i}[X \mid \mathbb{F}_t^Y]\big] \text{ since } 1_{A_k\cap A_i} \text{ is 0 except when } k = i.
\end{aligned}$$

Second, we deduce from

$$\mathrm{E}[X \mid \mathbb{F}_t^Y] = \mathrm{E}\left[\mathrm{E}[X \mid \mathbb{F}_t^Y \vee \sigma(Y_{t+1})] \mid \mathbb{F}_t^Y\right]$$

and (5.3) that

$$\mathrm{E}[X \mid \mathbb{F}_t^Y] = \sum_{i=1}^{m} \mathrm{E}[1_{A_i} \mid \mathbb{F}_t^Y]\, \mathrm{E}_{A_i}[X \mid \mathbb{F}_t^Y].$$

Applying this equality to the random variable $1_{A_k} X$, we obtain (5.4).

Note that if $\mathrm{P}(A_k) = 0$ for some k, then $\mathrm{E}_{A_k}[X \mid \mathbb{F}_t^Y] = 0$. Hence, in (5.3), each term $1_{A_k}\, \mathrm{E}_{A_k}[X \mid \mathbb{F}_t^Y]$ for which $\mathrm{P}(A_k) = 0$, vanishes. We also mention that $\mathrm{P}(A_k) > 0$ implies that $\mathrm{P}(A_k \mid \mathbb{F}_t^Y) = \mathrm{E}[1_{A_k} \mid \mathbb{F}_t^Y] > 0$ P a.s. Therefore, by (5.4), a nonzero term in (5.3) may be written as the following well-defined fraction (up to a set of P-probability zero)

$$\frac{\mathrm{E}[1_{A_k}\, X \mid \mathbb{F}_t^Y]}{\mathrm{E}[1_{A_k} \mid \mathbb{F}_t^Y]}.$$

Formula (5.5) is just a reformulation of (5.3) taking into account Relation (5.4) and the previous comments. This complete the proof. □

Proof[. Proof of Theorem 2.1] For $t = 0$, we deduce from (5.1) and (5.2) that

$$\begin{aligned}
\widehat{X}_0 := \mathrm{E}[X_0 \mid Y_0] &= \sum_{k=1}^{m} \left(\frac{\mathrm{E}[X_0 \langle Y_0, \boldsymbol{f}_k \rangle]}{\mathrm{E}[\langle Y_0, \boldsymbol{f}_k \rangle]} \right)^+ \langle Y_0, \boldsymbol{f}_k \rangle \\
&= \sum_{k=1}^{m} \left(\frac{\boldsymbol{p}_k}{\mathbf{1}^\mathsf{T} \boldsymbol{p}_k} \right)^+ \langle Y_0, \boldsymbol{f}_k \rangle
\end{aligned}$$

with the notation introduced in Theorem 2.1.

An application of Lemma 5.1 allows us to complete the proof of (2.10). Indeed, we find from our basic formula (5.5) that, for any $t \geq 0$,

$$\mathrm{E}\big[X_{t+1} \mid \mathbb{F}_t^Y \vee \sigma(Y_{t+1})\big] \;=\; \sum_{k=1}^{m} \left(\frac{\mathrm{E}\big[X_{t+1} \langle Y_{t+1}, \boldsymbol{f}_k \rangle \mid \mathbb{F}_t^Y\big]}{\mathrm{E}\big[\langle Y_{t+1}, \boldsymbol{f}_k \rangle \mid \mathbb{F}_t^Y\big]} \right)^+ \langle Y_{t+1}, \boldsymbol{f}_k \rangle.$$

The numerator of the fraction above may be rewritten as

$$\begin{aligned}
\mathrm{E}\big[X_{t+1} \langle Y_{t+1}, \boldsymbol{f}_k \rangle \mid \mathbb{F}_t^Y\big] &= \mathrm{E}\big[\mathrm{E}[X_{t+1} Y_{t+1}^\mathsf{T} \boldsymbol{f}_k \mid \mathbb{F}_t^{X,Y}] \mid \mathbb{F}_t^Y\big] \\
&= \mathrm{E}\big[\mathrm{E}[X_{t+1} {Y_{t+1}}^\mathsf{T} \mid \mathbb{F}_t^{X,Y}] \mid \mathbb{F}_t^Y\big] \boldsymbol{f}_k \\
&= \mathrm{E}\big[\boldsymbol{S}(t; X_t, Y_t) \mid \mathbb{F}_t^Y\big] \boldsymbol{f}_k \quad \text{from (2.6)} \\
&= \boldsymbol{S}(t; \widehat{X}_t, Y_t) \boldsymbol{f}_k \quad \text{with (2.9).}
\end{aligned}$$

Since $\mathbf{1}^\mathsf{T} X_{t+1} = 1$, the scalar $\mathrm{E}[\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid \mathbb{F}_t^Y]$ is $\mathbf{1}^\mathsf{T}\mathrm{E}\big[X_{t+1}\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid \mathbb{F}_t^Y\big]$, or by the previous result, $\mathbf{1}^\mathsf{T}\boldsymbol{S}(t; \widehat{X}_t, Y_t)\boldsymbol{f}_k$. From this, we recover (2.10).

We recall that the scalar $\mathscr{L}_t^{k,ji}$ is defined by

$$\mathscr{L}_t^{k,ji} = \sum_{l=1}^{t} \langle Y_l, \boldsymbol{f}_k\rangle\langle X_l, \boldsymbol{e}_j\rangle\langle X_{l-1}, \boldsymbol{e}_i\rangle$$

for $t \geq 1$. Thus, we find that

$$\mathscr{L}_{t+1}^{k,ji} = \mathscr{L}_t^{k,ji} + \langle Y_{t+1}, \boldsymbol{f}_k\rangle\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle X_t, \boldsymbol{e}_i\rangle.$$

The corresponding decomposition of vector $\mathscr{L}_{t+1}^{k,ji} X_{t+1}$ therefore allows us to write

$$\begin{aligned}
\widehat{\mathscr{L}^{k,ji}} X_{t+1} &:= \mathrm{E}\big[\mathscr{L}_{t+1}^{k,ji} X_{t+1} \mid \mathbb{F}_{t+1}^Y\big] \\
&= \mathrm{E}\big[\mathscr{L}_t^{k,ji} X_{t+1} \mid \mathbb{F}_{t+1}^Y\big] + \mathrm{E}\big[\langle Y_{t+1}, \boldsymbol{f}_k\rangle\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle X_t, \boldsymbol{e}_i\rangle\, X_{t+1} \mid \mathbb{F}_{t+1}^Y\big] \\
&= \mathrm{E}\big[\mathscr{L}_t^{k,ji} X_{t+1} \mid \mathbb{F}_{t+1}^Y\big] + \langle Y_{t+1}, \boldsymbol{f}_k\rangle \mathrm{E}\big[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle X_t, \boldsymbol{e}_i\rangle \mid \mathbb{F}_{t+1}^Y\big]\, \boldsymbol{e}_j. \quad (5.6)
\end{aligned}$$

Using (5.5), the first term in the right-hand side of the above equality has the form

$$\sum_{l=1}^{m} \left(\frac{\mathrm{E}\big[\mathscr{L}_t^{k,ji} X_{t+1}\langle Y_{t+1}, \boldsymbol{f}_l\rangle \mid \mathbb{F}_t^Y\big]}{\mathrm{E}\big[\langle Y_{t+1}, \boldsymbol{f}_l\rangle \mid \mathbb{F}_t^Y\big]} \right)^+ \langle Y_{t+1}, \boldsymbol{f}_l\rangle.$$

Let us develop the numerator of the above fraction. Since $\mathscr{L}_t^{k,ji}$ is $\mathbb{F}_t^{X,Y}$-measurable, we have

$$\begin{aligned}
& \mathrm{E}\big[\mathscr{L}_t^{k,ji} X_{t+1}\langle Y_{t+1}, \boldsymbol{f}_l\rangle \mid \mathbb{F}_t^Y\big] \\
= \;& \mathrm{E}\big[\mathscr{L}_t^{k,ji} \mathrm{E}[X_{t+1} Y_{t+1}^\mathsf{T} \boldsymbol{f}_l \mid \mathbb{F}_t^{X,Y}] \mid \mathbb{F}_t^Y\big] \\
= \;& \mathrm{E}\big[\mathscr{L}_t^{k,ji} \mathrm{E}[X_{t+1} Y_{t+1}^\mathsf{T} \mid X_t, Y_t] \mid \mathbb{F}_t^Y\big] \boldsymbol{f}_l \quad \text{since } (X, Y) \text{ is Markovian} \\
= \;& \mathrm{E}\big[\mathscr{L}_t^{k,ji} \boldsymbol{S}(t; X_t, Y_t) \mid \mathbb{F}_t^Y\big] \boldsymbol{f}_l \quad \text{from (2.6)} \\
= \;& \sum_{p=1}^{n} \mathrm{E}\big[\mathscr{L}_t^{k,ji} \langle X_t, \boldsymbol{e}_p\rangle \boldsymbol{S}(t; \boldsymbol{e}_p, Y_t) \mid \mathbb{F}_t^Y\big] \boldsymbol{f}_l \\
= \;& \sum_{p=1}^{n} \mathrm{E}\big[\mathscr{L}_t^{k,ji} \langle X_t, \boldsymbol{e}_p\rangle \mid \mathbb{F}_t^Y\big] \boldsymbol{S}(t; \boldsymbol{e}_p, Y_t) \boldsymbol{f}_l \\
= \;& \sum_{p=1}^{n} \langle \mathrm{E}\big[\mathscr{L}_t^{k,ji} X_t \mid \mathbb{F}_t^Y\big], \boldsymbol{e}_p\rangle \boldsymbol{S}(t; \boldsymbol{e}_p, Y_t) \boldsymbol{f}_l \\
= \;& \boldsymbol{S}(t; \widehat{\mathscr{L}^{k,ji}} X_t, Y_t) \boldsymbol{f}_l \quad \text{with (2.8).} \qquad (5.7)
\end{aligned}$$

We now deduce, from Formula (5.5), that the scalar $\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mathrm{E}[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle X_t, \boldsymbol{e}_i\rangle \mid \mathbb{F}_{t+1}^Y]$ is equal to

$$\left(\frac{\mathrm{E}\big[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle X_t, \boldsymbol{e}_i\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid \mathbb{F}_t^Y\big]}{\mathrm{E}\big[\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid \mathbb{F}_t^Y\big]}\right)^+ \langle Y_{t+1}, \boldsymbol{f}_k\rangle.$$

The scalar in the numerator of the above fraction is rewritten as follows:

$$\begin{aligned}
\mathrm{E}\big[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle X_t, \boldsymbol{e}_i\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid \mathbb{F}_t^Y\big] &= \mathrm{E}\big[\langle X_t, \boldsymbol{e}_i\rangle \mathrm{E}[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid \mathbb{F}_t^{X,Y}] \mid \mathbb{F}_t^Y\big] \\
&= \mathrm{E}\big[\langle X_t, \boldsymbol{e}_i\rangle \mathrm{E}[e_j^\mathsf{T} X_{t+1} Y_{t+1}^\mathsf{T} \boldsymbol{f}_k \mid X_t, Y_t] \mid \mathbb{F}_t^Y\big] \\
&= \boldsymbol{e}_j^\mathsf{T} \mathrm{E}\big[\langle X_t, \boldsymbol{e}_i\rangle \mathrm{E}[X_{t+1} {Y_{t+1}}^\mathsf{T} \mid X_t, Y_t] \mid \mathbb{F}_t^Y\big] \boldsymbol{f}_k \\
&= \boldsymbol{e}_j^\mathsf{T} \mathrm{E}\big[\langle X_t, \boldsymbol{e}_i\rangle\, \boldsymbol{S}(t; \boldsymbol{e}_i, Y_t) \mid \mathbb{F}_t^Y\big] \boldsymbol{f}_k \quad \text{from (2.6)} \\
&= \boldsymbol{e}_j^\mathsf{T} \boldsymbol{S}(t; \boldsymbol{e}_i, Y_t) \boldsymbol{f}_k\, \mathrm{E}\big[\langle X_t, \boldsymbol{e}_i\rangle \mid \mathbb{F}_t^Y\big] \\
&= \boldsymbol{e}_j^\mathsf{T} \boldsymbol{S}(t; \boldsymbol{e}_i, Y_t) \boldsymbol{f}_k\, \langle \widehat{X}_t, \boldsymbol{e}_i\rangle.
\end{aligned}$$

Since $\mathrm{E}\big[\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mid \mathbb{F}_t^Y\big]$ was found to be $\mathbf{1}^\mathsf{T} \boldsymbol{S}(t; \widehat{X}_t, Y_t) \boldsymbol{f}_k$, it follows that

$$\langle Y_{t+1}, \boldsymbol{f}_k\rangle \mathrm{E}\big[\langle X_{t+1}, \boldsymbol{e}_j\rangle\langle X_t, \boldsymbol{e}_i\rangle \mid \mathbb{F}_{t+1}^Y\big] = \left(\frac{\boldsymbol{e}_j^\mathsf{T} \boldsymbol{S}(t; \boldsymbol{e}_i, Y_t) \boldsymbol{f}_k}{\mathbf{1}^\mathsf{T} \boldsymbol{S}(t; \widehat{X}_t, Y_t) \boldsymbol{f}_k}\, \langle \widehat{X}_t, \boldsymbol{e}_i\rangle\right)^+ \langle Y_{t+1}, \boldsymbol{f}_k\rangle. \quad (5.8)$$

Therefore, the recursion (2.13) for $\widehat{\mathscr{L}^{k,ji}X}$ finally takes the form of (5.6) using (5.7) and (5.8).

A direct proof of (2.11) and (2.12) could be given using Lemma 5.1. But for the sake of brevity, we remark that they can be easily derived from (2.13), which is associated with the filter $\widehat{\mathscr{L}^{k,ji}X}$, due to the relations

$$\widehat{\mathcal{O}^{(i)}X}_{t+1} = \sum_{k=1}^{m} \sum_{j=1}^{n} \widehat{\mathscr{L}^{k,ji}X}_{t+1}$$

$$\widehat{N^{ji}X}_{t+1} = \sum_{k=1}^{m} \widehat{\mathscr{L}^{k,ji}X}_{t+1}.$$

□

6 Conclusion

In this note, we have derived normalized filters for various statistics of any partially observed discrete-time finite Markov chain in a recursive form. These results cover, in particular, the case of time-homogeneous dependent hidden Markov chains discussed in [7] and

[8]. We emphasize that only simple computation involving discrete-time Markov chains are required. An application to the parameter estimation for discrete-time batch Markovian arrival processes has been outlined. Models of this sort are widely used in applied probability, especially in queuing theory.

References

[1] Alfa, A.S. and Neuts, M. (1995). Modelling vehicular traffic using the discrete time Markovian arrival process. *Transportation Science* **29,** 109–117.

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

[3] Aström, K.J. (1965). Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.* **10,** 174–205.

[4] Blondia, C. (1993). A discrete-time batch Markovian arrival process as B-ISDN traffic model. *Belgian J. Oper. Res., Statist. and Computer Science* **32,** 3–23.

[5] Boucheron, S. and Gassiat, E. (2005). An information-theoretic perspective on order estimation. In *Inference in Hidden Markov Models.* ed. Cappé, O., Moulines, E., and Ryden, T. Springer.

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

[7] Elliott, R. (1994). Recursive estimation for hidden Markov models: a dependent case. *Stoch. Anal. Appl.* **13,** 437–460.

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

[9] Elliott, R.J. and Huang, H. (1994). How to count and guess well: Discrete adaptative filters. *Appl. Math. Optim.* **30,** 51–78.

[10] Ephraim, Y. and Merhav, N. (2002). Hidden Markov processes. *IEEE Trans. Informat. Theory* **48,** 1518–1569.

[11] Gassiat, E. and Boucheron S. (2003). Optimal error exponents for hidden Markov model order estimation. *IEEE Trans. Informat. Theory* **49,** 864–880.

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

[13] Lin, D. and Makis, V. (2003). Filters for a partially observable system subject to random failure. *Adv. in Appl. Probab.* **35,** 207–227.

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

[15] Rabiner, L.R. (1989). A tutorial on hidden Markov models and seleceted applications speech recognition. *Proceedings of the IEEE* **77,** 257–286.

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

[17] Rydén, T. (1997). Estimating the order of continuous phase-type distributions and Markov-modulated Poisson processes. *Stoch. Models* **13,** 417–433.

[18] Segall, A. (1976). Recursive estimation from discrete-time point process. *IEEE Trans. Informat. Theory* **22,** 422–431.