Available at WWW.MATHEMATICSWEB.ORG
POWERED BY SCIENCE @ DIRECT®

Statistics & Probability Letters 62 (2003) 189–201

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

On hidden Markov chains and finite stochastic systems

Peter Spreij*

Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, Netherlands

Received July 2002; received in revised form December 2002

Abstract

In this paper we study various properties of finite stochastic systems or hidden Markov chains as they are alternatively called. We discuss their construction following different approaches and we also derive recursive filtering formulas for the different systems that we consider. A key tool is a simple lemma on conditional expectations.

MSC: 60G42; 60J10; 93E11

Keywords: Markov chain; Hidden Markov chain; Recursive filtering; Stochastic system

1. Introduction

This paper deals with *hidden Markov chains* (HMCs), stochastic processes that are nowadays widely used in applied fields such as signal processing, communication systems and biology, see for instance the examples and references in Künsch et al. (1995). We consider the case where the observed process (denoted by Y) and the underlying chain (denoted by X) take on finitely many values. HMCs are such that probabilities of future events of X and Y given the past only depend on the current state of X. Typically this means that X satisfies the role of a *state process* as it is used in stochastic system theory. One of the aims of the present paper is to shed some more light on the relation between stochastic systems and HMCs. There are two slightly different definitions of stochastic systems, related by a time shift of the observed process. We will see that an HMC satisfies both definitions. HMCs satisfy different factorization and splitting properties of conditional probabilities of the bivariate process (X, Y). These properties are related to the different concepts of a stochastic system. We will also study for the different constructions the filtering and prediction

* Corresponding author. Tel.: +31-20-525-6070; fax: +31-20-525-5101.
E-mail address: spreij@science.uva.nl (P. Spreij).

0167-7152/03/$ - see front matter
doi:10.1016/S0167-7152(03)00011-7

problems and show that the solutions coincide if one deals with a HMC in the way we define it. The paper is organized as follows.

In Section 2 we describe the probabilistic behaviour of the joint process (X,Y) in more detail using the outer product of X and Y and by using properties of Kronecker products of matrices. We also introduce some properties that are typical for HMCs.

In Section 3 we study these properties and related ones in more detail. The convenient tool is a simple key lemma, that is presented in the appendix, on conditional expectations that involves a finitely generated σ-algebra. It is also shown that HMCs can be described by what in the engineering literature are called stochastic systems. In particular, it is shown that HMCs are characterized by being stochastic systems in two different senses. It is also shown how these two notions are interrelated.

In Section 4 we show how various filtering and prediction formulas are simple consequences of the same key lemma on conditional expectations.

2. Preliminaries and motivation

Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space on which all the random variables to be encountered below are defined. Consider the following familiar construction of what later will be called an HMC. We have a (stationary) Markov chain X with finite state space $S=\{1,\ldots,n\}$ and transition probabilities $a_{ij}=\mathbb{P}(X_{t+1}=i|X_t=j)$ with $t\in T=\{0,1,\ldots\}$, A is the matrix with elements a_{ij}. Furthermore, we have for each $i\in S$ and $t\in T$ a random variable $h_t(i)$ that all take values in $\{1,\ldots,m\}$. We assume that the n-vectors $(h_t(1),\ldots,h_t(n))$ form an independent identically distributed (iid) sequence that is also independent of X. We let $g_{ji}=\mathbb{P}(h_t(i)=j)$ and put these probabilities into an $m\times n$ matrix G. Finally, we define the process Y by $Y_t=h(X_t)$ for all t. By specifying the distribution of X_0 we thus completely determine the law of the bivariate process (X,Y). Note that the distribution of X_0, since it is a unit-vector valued random variable, is given by the vector $\mathbb{E}X_0$.

A convenient way to represent the distribution of the above process (X,Y) is obtained by changing the state space S into $E=\{e_1,\ldots,e_n\}$, the set of standard basis vectors of $\mathbb{R}^n$ and the space where Y takes its values into the set $F=\{f_1,\ldots,f_m\}$ of basis vectors of $\mathbb{R}^m$. With $H_t=[h_t(e_1),\ldots,h_t(e_n)]$ we then get $Y_t=H_tX_t$. Note that $\mathbb{E}H_t=G$. Define the filtration $\mathbb{F}=\{\mathscr{F}_t\}$ by $\mathscr{F}_t=\sigma\{X_0,\ldots,X_t,H_0,\ldots,H_t\}$. Then X and Y are adapted to this filtration. We will call X the *state* process and Y the *observation* or *output* process. Throughout the paper we assume that each state e_i is reachable from the starting state with positive probability. If this were not the case, this can always be accomplished by reducing the state space of X by taking basis vectors of a lower-dimensional Euclidean space. We also assume (without loss of generality) the nondegeneracy condition that none of the rows of G is zero.

By construction, the joint process (X,Y) is Markov with respect to $\mathbb{F}$. For completeness we give its transition probabilities, already present in Baum and Petrie (1966).

$$\mathbb{P}(X_t=e_i,Y_t=f_j|\mathscr{F}_{t-1})=g_{ji}e_i^{\top}AX_{t-1}. \tag{1}$$

Let us now represent the one-step transition probabilities of the joint chain (X,Y) in matrix form. To this end it is useful to work with the process Z that is obtained by $Z_t=Y_t\otimes X_t$ ($\otimes$ denotes Kronecker product). Note also the trivial relations $X_t=(\mathbf{1}_m^{\top}\otimes I_n)Z_t$ and $Y_t=(I_m\otimes\mathbf{1}_n^{\top})Z_t$, with I_m being the m-dimensional identity matrix and $\mathbf{1}_n$ the n-dimensional column vector with all its elements equal to one.

The transition probabilities of (X, Y) may then be represented by those of the Markov chain Z and vice versa. If we put the latter in a $nm \times nm$ matrix Q, then we have according to (1) that Q can be decomposed into m^2 blocks Q_{ij} that are equal to $\mathrm{diag}(G_{i.})A$, where $G_{i.}$ is the ith row of G. For a more compact formulation we introduce (like in Spreij (2001)) the following notation. Given an $m \times n$ matrix G let $\Delta(G)$ be the $nm \times n$ matrix defined by

$$\Delta(G) = \begin{bmatrix} \mathrm{diag}(G_{1.}) \\ \vdots \\ \mathrm{diag}(G_{m.}) \end{bmatrix}.$$

Using the notation $\Delta(G)$ we can now write

$$Q = \Delta(G)A(\mathbf{1}_m^\top \otimes I_n). \tag{2}$$

We thus get (see Spreij, 2001) $\mathbb{E}[Z_t|\mathscr{F}_{t-1}] = QZ_{t-1}$ and the initial distribution of Z, also a unit-vector valued random variable, is given by the vector $\mathbb{E}Z_0 = \Delta(G)\mathbb{E}X_0$. One also obtains the relations $\mathbb{E}[X_t|\mathscr{F}_{t-1}] = AX_{t-1}$ and $\mathbb{E}[Y_t|\mathscr{F}_{t-1}] = GX_{t-1}$.

As an alternative to looking at the bivariate process (X, Y) via the process Z as above, we also study the process W, again built from X and Y and defined by $W_t = Y_{t-1} \otimes X_t$ for $t \geqslant 1$. Along with this process we consider the filtration $\mathbb{G}$ of σ-algebras $\mathscr{G}_t := \sigma\{H_0, \dots, H_{t-1}, X_0, \dots, X_t\}$. Then W is $\mathbb{G}$-adapted and the $\mathscr{G}_t$ and the $\mathscr{F}_t$ are related by $\mathscr{F}_{t-1} \vee \sigma(X_t) = \mathscr{G}_t$ and $\mathscr{G}_t \vee \sigma(H_t) = \mathscr{F}_t$.

We obtain the relations

$$\mathbb{E}[W_t|\mathscr{F}_{t-1}] = (I_m \otimes A)Z_{t-1}, \tag{3}$$

$$\mathbb{E}[W_t|\mathscr{G}_{t-1}] = GX_{t-1} \otimes AX_{t-1} = (I_m \otimes A)\Delta(G)X_{t-1}. \tag{4}$$

In particular, it follows that W is $\mathbb{G}$-Markov with transition matrix

$$R := (I_m \otimes A)\Delta(G)(\mathbf{1}_m^\top \otimes I_n). \tag{5}$$

We mention some properties that (X, Y) has:

1. (X, Y) is a stochastic system in the sense of Picci (1978), since Z is $\mathbb{F}$-Markov (see Section 3).
2. It is immediate from the construction at the beginning of this section, or from Eq. (1), that what is called the "*factorization property*" (Finesso, 1990) holds:

$$\mathbb{P}(X_t = e_i, Y_t = f_j|\mathscr{F}_{t-1}) = \mathbb{P}(Y_t = f_j|X_t = e_i)\mathbb{P}(X_t = e_i|X_{t-1}), \tag{6}$$

   which can compactly be formulated in terms of Z as

$$\mathbb{E}[Z_t|\mathscr{F}_{t-1}] = \Delta(G)\mathbb{E}[X_t|X_{t-1}] \quad \forall t. \tag{7}$$

3. (X, Y) is a stochastic system in the sense of van Schuppen (1989), (see Section 3), since W is $\mathbb{G}$-Markov.
4. W has, as opposed to the factorization property, the *splitting property*

$$\mathbb{E}[W_{t+1}|\mathscr{G}_t] = \mathbb{E}[Y_t|\mathscr{G}_t] \otimes \mathbb{E}[X_{t+1}|\mathscr{G}_t], \tag{8}$$

   which immediately follows from (4).

In the subsequent section we will see how these (necessary) properties can be used to construct stochastic systems that result in HMCs.

Remark 2.1. All the properties mentioned above in terms of conditional expectations given the σ-algebras $\mathscr{F}_t$ and $\mathscr{G}_t$ remain valid if we replace these with $\sigma\{X_0,\dots,X_t,Y_0,\dots,Y_t\}$ and $\sigma\{X_0,\dots,X_t, Y_0,\dots,Y_{t-1}\}$. Hence the law of the bivariate process (X,Y), being a Markov chain with respect to its own filtration, is completely specified by the matrices A and G and the initial law of X. It follows that any bivariate Markov process (X,Y) can be generated by the mechanism $Y_t = H_tX_t$ of this section, if we assume the following: The transition matrix Q of the associated process Z defined by $Z_t = Y_t \otimes X_t$ is of form (2) and the initial law is given by $\mathbb{E}Z_0 = \Delta(G)\mathbb{E}X_0$.

In view of Remark 2.1 above we adopt the following

Definition 2.2. A bivariate process (X,Y) that assumes finitely many values is called an HMC if the process $Z = Y \otimes X$ is Markov with respect to its own filtration and if its matrix of transition probabilities is of the type given by (2).

Remark 2.3. The assumption in this section that the sequence $\{H_t\}$ is iid is equivalent to assuming that $\{H_t - G\}$ is a martingale difference sequence with respect to its own filtration. This is due to the fact that the columns of H_t are basis vectors and the argument is as follows. If $\{H_t - G\}$ is a martingale difference sequence, so is $\{h_t - g\}$, where $h_t = \mathrm{vec}(H_t)$ and $g = \mathrm{vec}(G)$. The h_t are again unit-vector valued and for a unit-vector e we thus have $\mathbb{P}(h_t = e|h_0,\dots,h_{t-1}) = \mathbb{P}(e^\top h_t = 1|h_0,\dots,h_{t-1}) = \mathbb{E}[e^\top h_t|h_0,\dots,h_{t-1}] = e^\top g$. Since this conditional probability does not depend on $h_0,\dots,h_{t-1}$, nor on t, we obtain that the sequence $\{h_t\}$, and hence also $\{H_t\}$, is iid. Hence the iid assumption is not as restrictive as it may seem at first glance.

3. HMCs and stochastic systems

There are various ways to describe properties of an HMC. We mention a few possibilities and show how these can be used as building stones for an HMC, together with notions from stochastic system theory.

Let X and Y be two stochastic processes taking values in the sets $\mathbb{E}$ and F, respectively, like in Section 2. Let Z again be the process $Y \otimes X$. For the time being, no further assumptions on X and Y are imposed, except that redundant states are excluded in the sense that each state of X is visited at least once with probability one and likewise for Y.

Throughout the rest of the paper we assume that for all t the σ-algebra $\mathscr{F}_t$ is generated by $X_0,\dots,X_t$ and $Y_0,\dots,Y_t$. The family $\{\mathscr{F}_t\}$ is again denoted by $\mathbb{F}$. We also consider the process W again, with $W_t = Y_{t-1} \otimes X_t$, adapted to the filtration $\mathbb{G} = \{\mathscr{G}_t\}$, with $\mathscr{G}_t$ generated by $X_0,\dots,X_t,Y_0,\dots,Y_{t-1}$. Note the relations

$$\mathscr{F}_t = \mathscr{G}_t \vee \sigma(Y_t),$$

$$\mathscr{G}_t = \mathscr{F}_{t-1} \vee \sigma(X_t).$$

In the previous section we restricted ourselves to time-invariant processes, implying that all conditional probabilities and expectations do not depend on time directly. Many properties of an HMC can be formulated without assuming stationarity. We introduce some notation. Given a stochastic process ζ with values in some arbitrary measurable space, we denote for all t by $\mathscr{F}_t^{\zeta}$ the σ-algebra generated by the ζ_s for $s \leqslant t$ and by $\mathscr{F}_t^{\zeta+}$ the σ-algebra generated by the ζ_s for $s \geqslant t$. Many of the results in the previous sections can be abstractly formulated in terms of properties of *stochastic systems*. A stochastic system is a formally defined concept. The main ingredients are a *state process* X and an *output process* Y (defined on a suitable probability space and taking values in some other spaces) and certain conditional independence relations.

Let us therefore recall some facts on conditional independence. Two σ-algebras $\mathscr{H}_1$ and $\mathscr{H}_2$ are called conditionally independent given a σ-algebra $\mathscr{G}$, if for all bounded $\mathscr{H}_i$-measurable functions H_i $(i=1,2)$, the relation $\mathbb{E}[H_1H_2|\mathscr{G}] = \mathbb{E}[H_1|\mathscr{G}]\mathbb{E}[H_2|\mathscr{G}]$ holds. A convenient characterization of this is that σ-algebras $\mathscr{H}_1$ and $\mathscr{H}_2$ are conditionally independent given σ-algebra $\mathscr{G}$, if for all bounded $\mathscr{H}_1$-measurable functions H_1, the relation $\mathbb{E}[H_1|\mathscr{G} \vee \mathscr{H}_2] = \mathbb{E}[H_1|\mathscr{G}]$ holds.

In the literature one can find two definitions of a stochastic system, that are slightly different. The first one is due to Picci (1978), and the essential part of the definition is that for all t the σ-algebras $\mathscr{F}_{t+1}^{X+} \vee \mathscr{F}_{t+1}^{Y+}$ and $\mathscr{F}_t^X \vee \mathscr{F}_t^Y$ are conditionally independent given $\sigma(X_t)$. The other one is due to van Schuppen (1989) in which the conditional independence relation between σ-algebras becomes: for all t the σ-algebras $\mathscr{F}_t^{X+} \vee \mathscr{F}_t^{Y+}$ and $\mathscr{F}_{t-1}^X \vee \mathscr{F}_{t-1}^Y$ are conditionally independent given $\sigma(X_t)$. Implications of the two different definitions for the filtering problem will be discussed in Section 4.

We will write $(X,Y) \in \Sigma_{\mathrm{P}}$ if the pair of processes (X,Y) is a stochastic system according to Picci (1978) and $(X,Y) \in \Sigma_{\mathrm{S}}$ if it is one in the sense of van Schuppen (1989). Using this notation, we see that $(X,Y) \in \Sigma_{\mathrm{P}}$ is equivalent with saying that Z is an $\mathbb{F}$-Markov process with transition probabilities depending on X only, and that $(X,Y) \in \Sigma_{\mathrm{S}}$ is equivalent with saying that W is a $\mathbb{G}$-Markov process with transition probabilities depending on X only. Note that both for stochastic system (X,Y) either in Σ_{P} or in Σ_{S} the state process is always Markov relative to its own filtration.

Let us consider a set of possible properties that the processes X and Y may possess, motivated by what we found in Section 2 for HMCs. We gather these properties in the next

Definition 3.1. Let X and Y be unit-vector valued processes.

1. We say that the *output property* holds if
$$\mathbb{E}[Y_t|\mathscr{G}_t] = \mathbb{E}[Y_t|\mathscr{F}_{t-1} \vee \sigma(X_t)] = \mathbb{E}[Y_t|\sigma(X_t)] \quad \forall t. \tag{9}$$
Alternatively, one can say that the output property holds if the sigma-algebras $\sigma(Y_t)$ and $\mathscr{F}_{t-1}$ are conditionally independent given $\sigma(X_t)$. If this property holds together with time invariance, we use the matrix G defined by $\mathbb{E}[Y_t|\sigma(X_t)] = GX_t$, where we also assume that G is not depending on t. G is then such that the columns $G_{\cdot i}$ are equal to $\mathbb{E}[Y_t|X_t = e_i]$.
2. The processes X and Y enjoy the *factorization property* if there exists a matrix $K \in \mathbb{R}^{m\times n}$ such that
$$\mathbb{E}[Z_t|\mathscr{F}_{t-1}] = \Delta(K)\mathbb{E}[X_t|\mathscr{F}_{t-1}] \quad \forall t. \tag{10}$$
3. The *splitting property* holds if
$$\mathbb{E}[W_{t+1}|\mathscr{G}_t] = \mathbb{E}[Y_t|\mathscr{G}_t] \otimes \mathbb{E}[X_{t+1}|\mathscr{G}_t] \quad \forall t. \tag{11}$$

Hence the splitting property says that $\sigma(X_{t+1})$ and $\sigma(Y_t)$ are conditionally independent given $\mathscr{G}_t$. This can alternatively expressed by $\mathbb{E}[X_{t+1}|\mathscr{F}_t] = \mathbb{E}[X_t|\mathscr{G}_t]$, because of the above given characterization of conditional independence and the relation $\mathscr{F}_t = \mathscr{G}_t \vee \sigma(Y_t)$.

First we comment on the factorization property. We showed that it is valid for the HMC of Section 2. But one can always factorize $\mathbb{E}[Z_t|\mathscr{F}_{t-1}]$ with a second factor $\mathbb{E}[X_t|\mathscr{F}_{t-1}]$ as in (10), however, in general, the left factor is a random ($\mathscr{F}_{t-1}$-measurable) diagonal matrix (see Eq. (14) below).

Denote by $\mathbb{P}_i$ the conditional measure on $(\Omega, \mathscr{F})$ given $X_t = e_i$. Expectation with respect to these measures will be denoted by $\mathbb{E}_i$, with the understanding that expectations $\mathbb{E}_i U$ are set equal to zero, if $\mathbb{P}(X_t = e_i) = 0$ (cf. the appendix). Then for any sub-σ-algebra $\mathscr{F}^0$ of $\mathscr{F}$ and any integrable random variable U we have from Eq. (A.3) in the appendix the relation

$$\mathbb{E}[U 1_{\{X_t = e_i\}}|\mathscr{F}^0] = \mathbb{E}_i[U|\mathscr{F}^0]\mathbb{P}(X_t = e_i|\mathscr{F}^0). \tag{12}$$

Application of Eq. (12) with $U = f_j^\top Y_t$, $\mathscr{F}^0 = \mathscr{F}_{t-1}$ for all i yields

$$\mathbb{E}[X_t f_j^\top Y_t|\mathscr{F}_{t-1}] = \mathrm{diag}\,(m_j)\mathbb{E}[X_t|\mathscr{F}_{t-1}], \tag{13}$$

where m_j is the column vector with ith element $\mathbb{E}_i[f_j^\top Y_t|\mathscr{F}_{t-1}]$. Let M be the matrix that has ith column $\mathbb{E}_i[Y_t|\mathscr{F}_{t-1}]$. Then we get

$$\mathbb{E}[Z_t|\mathscr{F}_{t-1}] = \Delta(M)\mathbb{E}[X_t|\mathscr{F}_{t-1}]. \tag{14}$$

Proposition 3.2. *Properties* 1 *and* 2 *of Definition* 3.1 *are equivalent if the processes are time invariant. Moreover, in that case one has* $K = G$.

Proof. Assume that the output property holds. Use then reconditioning in (9) to get: $\mathbb{E}[Z_t|\mathscr{F}_{t-1}] = \mathbb{E}[\mathbb{E}[Z_t|\mathscr{F}_{t-1} \vee \sigma(X_t)]|\mathscr{F}_{t-1}] = \mathbb{E}[\mathbb{E}[Y_t|\mathscr{G}_t] \otimes X_t|\mathscr{F}_{t-1}] = \mathbb{E}[\mathbb{E}[Y_t|X_t] \otimes X_t|\mathscr{F}_{t-1}] = \mathbb{E}[GX_t \otimes X_t|\mathscr{F}_{t-1}] = \mathbb{E}[\Delta(G)X_t|\mathscr{F}_{t-1}]$. It follows from (14) that $\Delta(G) = \Delta(M)$, but since G is nonrandom, the validity of the factorization property follows.

Conversely, assume that the factorization property 2 holds. Take expectations in (10). Then $\mathbb{E}Z_t = \Delta(K)\mathbb{E}X_t$. From the definition of G (in property 1) we get $\mathbb{E}Z_t = \mathbb{E}\mathbb{E}[Y_t|\sigma(X_t) \otimes X_t] = \mathbb{E}(GX_t \otimes X_t) = \Delta(G)\mathbb{E}X_t$. Since for each i there is a t such that the ith component of $\mathbb{E}X_t$ is strictly positive, it follows from the blockwise diagonal structure of the Δ-matrices that $\Delta(G) = \Delta(K)$ and $G = K$.

Next we show that the output property holds. Assume for a moment that all elements of $\mathbb{E}X_t$ are positive. According to Eqs. (A.2) and (A.3) we have

$$\mathbb{E}[Y_t|\mathscr{F}_{t-1} \vee \sigma(X_t)] = \sum_i \frac{\mathbb{E}[Y_t e_i^\top X_t|\mathscr{F}_{t-1}]}{e_i^\top \mathbb{E}[X_t|\mathscr{F}_{t-1}]} e_i^\top X_t.$$

Since $Y_t e_i^\top X_t = (I_m \otimes e_i^\top)Z_t$ and using the factorization property, we can rewrite this as

$$\sum_i \frac{(I_m \otimes e_i^\top)\Delta(G)\mathbb{E}[X_t|\mathscr{F}_{t-1}]}{e_i^\top \mathbb{E}[X_t|\mathscr{F}_{t-1}]} e_i^\top X_t.$$

Because $(I_m \otimes e_i^\top)\Delta(G) = Ge_ie_i^\top$, this reduces to

$$G\sum_i \frac{e_ie_i^\top \mathbb{E}[X_t|\mathscr{F}_{t-1}]}{e_i^\top \mathbb{E}[X_t|\mathscr{F}_{t-1}]} e_i^\top X_t,$$

which in turn is nothing else but GX_t, from which we obtain the output property. In the case where the vector $\mathbb{E}X_t$ has some elements equal to zero, the above procedure is still valid, provided we let the summation indices run through the set $\{i : e_i^\top \mathbb{E}X_t > 0\}$. □

An obvious relation between the different concepts of stochastic system is that $(X,Y)\in\Sigma_{\mathrm{P}}$ iff $(X,\sigma Y)\in\Sigma_{\mathrm{S}}$, where σY is the process defined by $\sigma Y_t = Y_{t+1}$. Another relation, involving the splitting and output properties, is given in the following

Proposition 3.3. *A pair (X,Y) belongs to Σ_{S} and the splitting property holds iff it belongs to Σ_{P} and the output property (or the factorization property) holds.*

Proof. Suppose that $(X,Y)\in\Sigma_{\mathrm{P}}$ and that the output property holds. Since X is $\mathbb{F}$-Markov, we have $\mathbb{E}[X_{t+1}|\mathscr{F}_t]=\mathbb{E}[X_{t+1}|X_t]$, which is $\mathscr{G}_t$ measurable and therefore equal to $\mathbb{E}[X_{t+1}|\mathscr{G}_t]$, which is equivalent to the splitting property.

Next we show that (X,Y) also belongs to Σ_{S}. We compute

$$\begin{aligned}\mathbb{E}[W_{t+1}|\mathscr{G}_t] &= \mathbb{E}[\mathbb{E}[W_{t+1}|\mathscr{F}_t]|\mathscr{G}_t]\\ &= \mathbb{E}[Y_t\otimes \mathbb{E}[X_{t+1}|\mathscr{F}_t]|\mathscr{G}_t]\\ &= \mathbb{E}[Y_t\otimes \mathbb{E}[X_{t+1}|X_t]|\mathscr{G}_t]\\ &= \mathbb{E}[Y_t|\mathscr{G}_t]\otimes \mathbb{E}[X_{t+1}|X_t],\end{aligned}$$

which is $\sigma(X_t)$-measurable because of the output property.

Conversely, letting $(X,Y)\in\Sigma_{\mathrm{S}}$ we automatically get the output property, because $\mathbb{E}[Y_t|\mathscr{G}_t]=(I_n\otimes \mathbf{1}_n^\top)\mathbb{E}[W_{t+1}|\mathscr{G}_t]=(I_n\otimes \mathbf{1}_n^\top)\mathbb{E}[W_{t+1}|X_t]$ in view of $(X,Y)\in\Sigma_{\mathrm{S}}$. Assuming the conditional independence relation we obtain the Markov property of Z from

$$\begin{aligned}\mathbb{E}[Z_{t+1}|\mathscr{F}_t] &= \mathbb{E}[\mathbb{E}[Z_{t+1}|\mathscr{G}_{t+1}]|\mathscr{F}_t]\\ &= \mathbb{E}[\mathbb{E}[Y_{t+1}|\mathscr{G}_{t+1}]\otimes X_{t+1}|\mathscr{F}_t]\\ &= \mathbb{E}[\mathbb{E}[Y_{t+1}|X_{t+1}]\otimes X_{t+1}|\mathscr{F}_t] \quad \text{(output property)}\\ &= \mathbb{E}[\mathbb{E}[Y_{t+1}|X_{t+1}]\otimes X_{t+1}|\mathscr{G}_t] \quad \text{(splitting property)}\\ &= \mathbb{E}[\mathbb{E}[Y_{t+1}|X_{t+1}]\otimes X_{t+1}|X_t] \quad (W \text{ is } \mathbb{G}\text{-Markov}),\end{aligned}$$

which shows that $(X,Y)\in\Sigma_{\mathrm{P}}$. □

The connection between systems in Σ_{P} and Σ_{S} and HMCs can be characterized as follows.

Proposition 3.4. *A finite valued time-invariant system belonging both to Σ_{P} and to Σ_{S} is an HMC and vice versa.*

Proof. We have seen already in Section 2 that an HMC belongs to both $\Sigma_{\rm P}$ and to $\Sigma_{\rm S}$. Conversely, let a pair (X,Y) belong to these two classes. Since it belongs to $\Sigma_{\rm S}$ we conclude from the proof of Proposition 3.3 that the output property holds and thus in view of Proposition 3.2 also the factorization property, saying that $\mathbb{E}[Z_t|\mathscr{F}_{t-1}] = \Delta(G)\mathbb{E}[X_t|\mathscr{F}_{t-1}]$. Since the process X is Markov w.r.t. $\mathbb{F}$, with transition matrix A say, we get $\mathbb{E}[Z_t|\mathscr{F}_{t-1}] = \Delta(G)AX_{t-1} = \Delta(G)A(\mathbf{1}^\top \otimes I_n)Z_{t-1}$. Therefore, the result follows from Definition 2.2. □

4. Filtering and prediction

In this section we give some filtering and prediction formulas. By the filtering problem for a system (X,Y) belonging to $\Sigma_{\rm P}$ or to $\Sigma_{\rm S}$ we mean the determination for each t of the conditional law of X_t given $Y_0,\dots,Y_t$. As before, for each t we denote by $\mathscr{F}^Y_t$ the σ-algebra generated by $Y_0,\dots,Y_t$. Since the state space of X is a set of basis vectors, this conditional law is completely determined by the conditional expectation $\mathbb{E}[X_t|\mathscr{F}^Y_t]$. The prediction problem is to determine for each t the conditional law of X_{t+1} given $Y_0,\dots,Y_t$, that is completely characterized by the conditional expectations $\mathbb{E}[X_{t+1}|\mathscr{F}^Y_t]$. We will use the notations $\mathbb{E}[X_t|\mathscr{F}^Y_t]=\hat{X}_t$ and $\mathbb{E}[X_{t+1}|\mathscr{F}^Y_t]=\hat{X}_{t+1|t}$. Similarly, we write $\mathbb{E}[Y_{t+1}|\mathscr{F}^Y_t]=\hat{Y}_{t+1|t}$. In addition to the above, one wants to have $\hat{X}_t$ and $\hat{X}_{t+1|t}$ in recursive form. We shall see below that the recursions for the cases $(X,Y)\in\Sigma_{\rm P}$ and $(X,Y)\in\Sigma_{\rm S}$ are different.

In the book (Elliott et al., 1995) recursive formulae for unnormalized filters are obtained by a measure transformation. Here we undertake a direct approach that leads to a simple recursive formula for the conditional probabilities itself. The key argument in all cases is provided by Lemma A.1.

4.1. Filter for $\Sigma_{\rm P}$

In this section we obtain the filter for a system in $\Sigma_{\rm P}$, so we work with a Markov chain $Z_t=X_t\otimes Y_t$ with transition matrix $Q=\bar{Q}(\mathbf{1}_m^\top\otimes I_n)$. We can write the matrix $\bar{Q}$ as

$$\bar{Q} = \begin{bmatrix} Q_1 \\ \vdots \\ Q_m \end{bmatrix} \tag{15}$$

with the Q_i in $\mathbb{R}^{n\times n}$. No further assumptions on the Q_i are made. Observe that the Q_i have the interpretation that

$$Q_iX_t = \mathbb{E}[X_{t+1}1_{\{Y_{t+1}=f_i\}}|\mathscr{F}_t]. \tag{16}$$

We have the following result (alternatively presented in Picci (1978)).

Theorem 4.1. *The filter $\hat{X}$ is given by the recursion*

$$\hat{X}_t = \left[\frac{Q_1\hat{X}_{t-1}}{\mathbf{1}_n^\top Q_1\hat{X}_{t-1}} \quad \cdots \quad \frac{Q_m\hat{X}_{t-1}}{\mathbf{1}_n^\top Q_m\hat{X}_{t-1}}\right] Y_t \tag{17}$$

with the initial condition determined by the initial law of Z. *The prediction* $\hat{X}_{t+1|t}$ *is equal to* $A\hat{X}_t$ *with* $A = \sum_{i=1}^m Q_i$ *and* $X_{0|-1} = \mathbb{E}X_0 = p_0$. *For the prediction* $\hat{Y}_{t+1|t}$ *we have* $\hat{Y}_{t+1|t} = C\hat{X}_t$ *with* $C = (I_m \otimes \mathbf{1}_n^\top)\bar{Q}$.

Proof. We use Eq. (A.2) with $\mathscr{F}^0 = \mathscr{F}_t^Y$, $\mathscr{H} = \sigma(Y_{t+1})$, which is generated by the sets $H_i = \{Y_{t+1} = f_i\}$ and $U = X_{t+1}$. Thus, we obtain

$$\mathbb{E}[X_{t+1}|\mathscr{F}_{t+1}^Y] = \sum_{i=1}^m \mathbb{E}_i[X_{t+1}|\mathscr{F}_t^Y]1_{H_i} = \sum_{i=1}^m \frac{\mathbb{E}[X_{t+1}1_{H_i}|\mathscr{F}_t^Y]}{\mathbb{P}(H_i|\mathscr{F}_t^Y)}1_{H_i}.$$

Then we use the Markov property of Z to write

$$\mathbb{E}[X_{t+1}1_{H_i}|\mathscr{F}_t^Y] = \mathbb{E}[\mathbb{E}[X_{t+1}1_{H_i}|\mathscr{F}_t]|\mathscr{F}_t^Y] = \mathbb{E}[Q_iX_t|\mathscr{F}_t^Y] = Q_i\hat{X}_t.$$

Since $\mathbb{P}(H_i|\mathscr{F}_t^Y) = \mathbb{E}[1_{H_i}|\mathscr{F}_t] = \mathbf{1}_n^\top \mathbb{E}[X_{t+1}1_{H_i}|\mathscr{F}_t^Y]$ we get Eq. (17).

Define now $A = \sum_{i=1}^m Q_i = (\mathbf{1}_m^\top \otimes I_n)\bar{Q}$ and $C = (I_m \otimes \mathbf{1}_n^\top)\bar{Q}$. Then we have $\mathbb{E}[X_{t+1}|\mathscr{F}_t] = AX_t$ and $\mathbb{E}[Y_{t+1}|\mathscr{F}_t] = CX_t$. As a consequence we get by reconditioning that $\hat{X}_{t+1|t} = A\hat{X}_t$ and that $\hat{Y}_{t+1|t} = C\hat{X}_t$. □

We see that the filter $\hat{X}_t$ satisfies a completely recursive system, that is, $\hat{X}_t$ is completely determined by $\hat{X}_{t-1}$ and Y_t. In the absence of further conditions on the matrix Q (in particular the factorization property), there seems to be no complete recursion that is satisfied by $X_{t|t-1}$. The reason for this is that we did not have the Markov property of W with respect to $\mathbb{G}$, unless the factorization property holds, in which case the formulas above take a particular nice form (see Section 4.3).

Remark 4.2. It follows from Eq. (16) that filter (17) can alternatively be expressed as

$$\hat{X}_t = \left[Q_1\hat{X}_{t-1}, \ldots, Q_m\hat{X}_{t-1}\right] \operatorname{diag}(C\hat{X}_{t-1})^{-1}Y_t.$$

Indeed, from Eq. (16) we obtain $\mathbf{1}^\top Q_iX_t = \mathbb{P}(Y_{t+1} = f_i|\mathscr{F}_t)$, hence $\mathbb{E}[Y_{t+1}|\mathscr{F}_t]$ is the vector with elements $\mathbf{1}^\top Q_iX_t$. Conditioning of this vector on $\mathscr{F}_t^Y$ gives that $\hat{Y}_{t+1|t}$ is the vector with elements $\mathbf{1}^\top Q_i\hat{X}_t$. So we can rewrite (17) as $\hat{X}_t = [Q_1\hat{X}_{t-1}, \ldots, Q_m\hat{X}_{t-1}] \operatorname{diag}(Y_{t|t-1})^{-1}Y_t$ and the result follows.

4.2. Filter for Σ_S

In this section we obtain the filter for a system in Σ_S, so we work with a Markov chain $W_t = X_t \otimes Y_{t-1}$ with transition matrix $R = \bar{R}(\mathbf{1}_m^\top \otimes I_n)$, where the matrix $\bar{R}$ can be written as

$$\bar{R} = \begin{bmatrix} R_1 \\ \vdots \\ R_m \end{bmatrix} \tag{18}$$

for certain matrices R_i in $\mathbb{R}^{n\times n}$. No further assumptions on the R_i are made. Observe that the R_i have the interpretation that

$$R_iX_t = \mathbb{E}[X_{t+1}1_{\{Y_t = f_i\}}|\mathscr{G}_t]. \tag{19}$$

Then we have

Theorem 4.3. *The predictor* $\hat{X}_{t|t-1}$ *is given by the recursion*

$$\hat{X}_{t+1|t} = \left[\frac{R_1\hat{X}_{t|t-1}}{\mathbf{1}_n^\top R_1\hat{X}_{t|t-1}} \quad \cdots \quad \frac{R_m\hat{X}_{t|t-1}}{\mathbf{1}_n^\top R_m\hat{X}_{t|t-1}} \right] Y_t \tag{20}$$

with the initial condition $X_{0|-1}=\mathbb{E}X_0$. *For the filter* $\hat{X}_t$ *and for* $\hat{Y}_{t+1|t}$ *we have the following relations*:

$$\hat{X}_{t+1} = \mathrm{diag}(\hat{X}_{t+1|t})G^\top \mathrm{diag}(\hat{Y}_{t+1|t})^{-1}Y_{t+1}, \tag{21}$$

where $G=(I_m \otimes \mathbf{1}_n^\top)\bar{R}$ *and*

$$\hat{Y}_{t+1|t} = GX_{t+1|t}. \tag{22}$$

Proof. We use Eq. (A.2) with $\mathscr{F}^0=\mathscr{F}^Y_{t-1}$, $\mathscr{H}=\sigma(Y_t)$, which is generated by the sets $H_i=\{Y_t=f_i\}$ and $U=X_{t+1}$. Then we obtain

$$\mathbb{E}[X_{t+1}|\mathscr{F}^Y_t] = \sum_{i=1}^m \mathbb{E}_i[X_{t+1}|\mathscr{F}^Y_{t-1}]1_{H_i} = \sum_{i=1}^m \frac{\mathbb{E}[X_{t+1}1_{H_i}|\mathscr{F}^Y_{t-1}]}{\mathbb{P}(H_i|\mathscr{F}^Y_t)}1_{H_i}.$$

Then we use the Markov property of W to write

$$\begin{aligned}\mathbb{E}[X_{t+1}1_{H_i}|\mathscr{F}^Y_{t-1}] &= \mathbb{E}[\mathbb{E}[X_{t+1}1_{H_i}|\mathscr{G}_t]|\mathscr{F}^Y_{t-1}]\\ &= \mathbb{E}[R_iX_t|\mathscr{F}^Y_{t-1}]\\ &= R_i\hat{X}_{t|t-1}.\end{aligned}$$

Since $\mathbb{P}(H_i|\mathscr{F}^Y_{t-1}) = \mathbb{E}[1_{H_i}|\mathscr{F}^Y_{t-1}] = \mathbf{1}_n^\top \mathbb{E}[X_{t+1}1_{H_i}|\mathscr{F}^Y_{t-1}]$ we get Eq. (20).

To derive formula (21) for the filter we proceed similarly, using Lemma A.1 again with $U=X_{t+1}$, $\mathscr{F}^0=\mathscr{F}^Y_t$ and $\mathscr{H}=\sigma(Y_{t+1})$ generated by the sets $H_i=\{Y_{t+1}=f_i\}$. Then we can write Eq. (A.2) as $\mathbb{E}[X_{t+1}|\mathscr{F}^Y_{t+1}] = \mathbb{E}[X_{t+1}Y_{t+1}^\top|\mathscr{F}^Y_t]\,\mathrm{diag}(\hat{Y}_{t+1|t})^{-1}Y_{t+1}$,

$$\begin{aligned}\mathbb{E}[X_{t+1}Y_{t+1}^\top|\mathscr{F}^Y_t] &= \mathbb{E}[\mathbb{E}[X_{t+1}Y_{t+1}^\top|\mathscr{G}_{t+1}]|\mathscr{F}^Y_t]\\ &= \mathbb{E}[X_{t+1}\mathbb{E}[Y_{t+1}^\top|\mathscr{G}_{t+1}]|\mathscr{F}^Y_t]\\ &= \mathbb{E}[X_{t+1}(GX_{t+1})^\top|\mathscr{F}^Y_t]\\ &= \mathbb{E}[\mathrm{diag}(X_{t+1})|\mathscr{F}^Y_t]G^\top.\end{aligned}$$

Then Eq. (21) follows, as well as Eq. (22), since we have $\mathbb{E}[Y_{t+1}|\mathscr{F}^Y_t] = \mathbb{E}[Y_{t+1}X_{t+1}^\top|\mathscr{F}^Y_t]\mathbf{1}_n = G\,\mathrm{diag}(\hat{X}_{t+1|t})\mathbf{1}_n = G\hat{X}_{t+1|t}$. □

Remark 4.4. By a similar argument as in Remark 4.2 we can rewrite recursion (20) for the predictor as

$$\hat{X}_{t+1|t} = [R_1\hat{X}_{t|t-1},\dots,R_m\hat{X}_{t|t-1}]\,\mathrm{diag}(G\hat{X}_{t|t-1})^{-1}Y_t.$$

Remark 4.5. Note that in contrast with what we got in Section 4.1 for Σ_{P} here the *predictor* satisfies a completely recursive system, whereas we obtain the filter in terms of the predictor.

The formulas above take a particular nice form if the system satisfies the splitting property (see Section 4.3).

4.3. Filter for an HMC

In this section we give the recursive filtering formula for the stochastic system with the HMC Y as its output. Therefore, we can apply the results of Section 4.1 with the specification that $\bar{Q} = \Delta(G)A$, so we have $Q_i = \operatorname{diag}(G_{i.})A$ and $\mathbf{1}_n^\top Q_i = G_{i.}A$. The following holds.

Theorem 4.6. (i) *The conditional distribution of the X_t given $Y_0, \ldots, Y_t$ is recursively determined by*

$$\hat{X}_t = \operatorname{diag}(A\hat{X}_{t-1})G^\top \operatorname{diag}(GA\hat{X}_{t-1})^{-1} Y_t, \tag{23}$$

with initial condition $\hat{X}_0 = \operatorname{diag}(p_0)G^\top \operatorname{diag}(Gp_0)^{-1} Y_0$, *with* $p_0 = \mathbb{E}X_0$.

(ii) *The conditional distribution of the X_t given $Y_0, \ldots, Y_{t-1}$ is recursively determined by*

$$\hat{X}_{t+1|t} = A \operatorname{diag}(\hat{X}_{t|t-1})G^\top \operatorname{diag}(G\hat{X}_{t|t-1})^{-1} Y_t, \tag{24}$$

with initial condition $X_{0|-1} = \mathbb{E}X_0 = p_0$.

(iii) *The conditional expectation* $\hat{Y}_{t+1|t} = \mathbb{E}[Y_{t+1}|\mathscr{F}_t^Y]$ *is given by*

$$\hat{Y}_{t+1|t} = GA \operatorname{diag}(\hat{X}_{t|t-1})G^\top \operatorname{diag}(G\hat{X}_{t|t-1})^{-1} Y_t. \tag{25}$$

Proof. (i) Just use Eq. (17) and note that

$$\frac{Q_i\hat{X}_{t-1}}{\mathbf{1}_n^\top Q_i\hat{X}_{t-1}} = \operatorname{diag}(A\hat{X}_{t-1})\,\frac{G_{i.}^\top}{(GA\hat{X}_{t-1})_i}.$$

(ii) follows from (i), since we know from Theorem 4.1 that $\hat{X}_{t+1|t} = A\hat{X}_t$.

(iii) also follows from Theorem 4.1, upon verifying that C now becomes GA. □

Remark 4.7. Here both the filter and the predictor satisfy a complete recursive system. This is not surprising, because an HMC is a stochastic system belonging to both Σ_{P} and Σ_{S}. Note that Theorem 4.6 can alternatively be derived from Theorem 4.3, since under the assumptions of the present subsection we have that $R_i = A \operatorname{diag}(G_{i.})$.

Remark 4.8. If we define for $x \in \mathbb{R}_+^n$ the matrix

$$G_x := \operatorname{diag}(x)G^\top \operatorname{diag}(Gx)^{-1},$$

then Eqs. (23)–(25) take the form $\hat{X}_t = G_{A\hat{X}_{t-1}} Y_t$, $\hat{X}_{t+1|t} = AG_{\hat{X}_{t|t-1}} Y_t$ and $\hat{Y}_{t+1|t} = GAG_{\hat{X}_{t|t-1}} Y_t$.

One may check that under the condition that Y is a *deterministic* function of X (in which case the columns of G are basis vectors of $\mathbb{R}^m$) the matrices G_x are right pseudo-inverses of G.

Acknowledgements

The detailed and very constructive comments by a referee resulted in a substantial improvement of the paper.

Appendix A. A lemma on conditional expectations

Consider some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and let $\mathscr{H}$ be a sub-σ-algebra of $\mathscr{F}$ that is generated by a finite partition $\{H_1, \ldots, H_k\}$ of Ω, satisfying $\mathbb{P}(H_i) > 0$ for all i. We introduce the (conditional) probability measures $\mathbb{P}_i$ on $(\Omega, \mathscr{F})$ defined by $\mathbb{P}_i(F) = \mathbb{E}[1_F(1_{H_i}/\mathbb{P}(H_i))] = \mathbb{P}(F|H_i)$. Expectation with respect to $\mathbb{P}_i$ is denoted by $\mathbb{E}_i$. Recall that for any integrable random variable U it holds that

$$\mathbb{E}[U|\mathscr{H}] = \sum_{i=1}^{k} \mathbb{E}_i[U]1_{H_i}. \tag{A.1}$$

We extend this result in the following easy to prove lemma. It is used frequently in Sections 3 and 4.

Lemma A.1. *Let $\mathscr{F}^0$ be some sub-σ-algebra of $\mathscr{F}$. Then the following equalities hold true.*

$$\mathbb{E}[U|\mathscr{F}^0 \vee \mathscr{H}] = \sum_{i=1}^{k} \mathbb{E}_i[U|\mathscr{F}^0]1_{H_i}, \tag{A.2}$$

$$\mathbb{E}[1_{H_j}U|\mathscr{F}^0] = \mathbb{E}[1_{H_j}|\mathscr{F}^0]\mathbb{E}_j[U|\mathscr{F}^0]. \tag{A.3}$$

Proof. Concerning the first equality we have to show that

$$\mathbb{E}\{1_{F\cap H_j}U\} = \mathbb{E}\left\{1_{F\cap H_j} \sum_{i=1}^{k} \mathbb{E}_i[U|\mathscr{F}^0]1_{H_i}\right\}$$

for all $F \in \mathscr{F}^0$, because every set in $\mathscr{F}^0 \vee \mathscr{H}$ can be written as a finite union of sets $F \cap H_j$ with some $F \in \mathscr{F}^0$ and because the RHS of (A.2) is clearly $\mathscr{F}^0 \vee \mathscr{H}$-measurable. We develop

$$\begin{aligned}
\mathbb{E}\left\{1_{F\cap H_j} \sum_{i=1}^{k} \mathbb{E}_i[U|\mathscr{F}^0]1_{H_i}\right\} &= \mathbb{E}\{1_{F\cap H_j}\mathbb{E}_j[U|\mathscr{F}^0]\} \\
&= \mathbb{E}_j\{1_F\mathbb{E}_j[U|\mathscr{F}^0]\}\mathbb{P}(H_j) \\
&= \mathbb{E}_j\{1_F U\}\mathbb{P}(H_j) \\
&= \mathbb{E}\{1_{F\cap H_j}U\}.
\end{aligned}$$

In these computations we used the trivial identity $\mathbb{E}1_{H_i}X = \mathbb{P}(H_i)\mathbb{E}_i X$ in the second and fourth equality and the defining property of conditional expectation in the third. This proves (A.2).

The second equality is a direct consequence of the first by conditioning on $\mathscr{F}^0$. Equality (A3) follows from (A2) by taking $1_{H_j}U$ instead of U and by conditioning on $\mathscr{F}^0$. $\square$

Remark A.2. If we take for $\mathscr{F}^0$ the trivial σ-algebra, then (A.2) reduces to (A.1). If $\mathbb{P}(H_i)=0$ for some i, then $\mathbb{P}_i$ is not well defined but (A.2) is still valid provided we define $\mathbb{E}_i[U|\mathscr{F}^0]$ to be zero for such an i.

Eq. (A.3) is also known as the conditional Bayes theorem (cf. Elliott et al., 1995, p. 23).

References

Baum, L.E., Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Statist. 37, 1554–1563.

Elliott, R.J., Aggoun, L., Moore, J.B., 1995. Hidden Markov Models. Estimation and Control. Springer, Berlin.

Finesso, L., 1990. Consistent estimation of the order for Markov and hidden Markov chains, Dissertation, University of Maryland.

Künsch, H., Geman, S., Kehagias, A., 1995. Ann. Appl. Probab. 5 (3), 577–602.

Picci, G., 1978. On the internal structure of finite state stochastic processes. In: Mohler, R.R., Ruberti, A. (Eds.), Recent Developments in Variable Structure Systems, Springer Lecture Notes in Economics and Mathematical Systems, Vol. 162. Springer, Berlin, pp. 288–304.

van Schuppen, J.H., 1989. Stochastic realization problems. In: Nijmeijer, H., Schumacher, J.M. (Eds.), Three Decades of Mathematical System Theory, Springer Lecture Notes in Control and Information Sciences, Vol. 135. Springer, Berlin, pp. 480–523.

Spreij, P.J.C., 2001. On the Markov property of a hidden Markov chain. Statist. Probab. Lett. 52/3, 279–288.