ELSEVIER

Statistics & Probability Letters 52 (2001) 279–288

STATISTICS & PROBABILITY LETTERS

www.elsevier.nl/locate/stapro

On the Markov property of a finite hidden Markov chain

Peter Spreij[1]

Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

Received January 1999

Abstract

In this paper we study the question of the conditions under which a hidden Markov chain itself exhibits Markovian behaviour. An insightful method to answer this question is based on a recursive filtering formula for the underlying chain. © 2001 Elsevier Science B.V. All rights reserved

MSC: 60J10; 93E11

Keywords: Markov chain; Hidden Markov chain; Recursive filtering

1. Introduction and preliminaries

The question whether a (deterministic) function of a Markov chain inherits the Markov property, has received much attention in the literature over the past years.See Section 2 for a short discussion on this. In this paper we pose the same question for a random function of a Markov chain. We provide an answer based on two different approaches. In Section 2, we use known results by Rubino and Sericola for deterministic functions of a Markov chain by considering a bivariate chain. In Section 3, we present a solution based on recursive filtering that is easier to interpret. We continue this section by introducing some notation and presenting some preliminary results.

Let $(\Omega, \mathscr{F}, P)$ be a probability space on which all the random variables to be encountered below are defined. Consider the following model for a hidden Markov chain (HMC):

$$X_t = AX_{t-1} + \varepsilon_t, X_0 \tag{1.1}$$

$$Y_t = H_t X_t \tag{1.2}$$

Here the *state* process X is modelled as a Markov process on the set $E = \{e_1, \dots, e_n\}$ of basis vectors of $\mathbb{R}^n$. Moreover, this process is supposed to be time-homogeneous with A the matrix of one step transition

E-mail address: spreij@science.uva.nl (P. Spreij).

[1] Tel.: +31-20-525-6070; fax: +31-20-525-5101

PII: S0167-7152(00)00216-9

probabilities: $A_{ij} = P(X_{t+1} = e_i | X_t = e_j)$. The process $\{\varepsilon_t\}$ is then a martingale difference sequence adapted to the filtration generated by X. This way of representing a finite Markov chain is common in engineering literature (see Elliott et al., 1995, p. 17). For Markov processes where the time set is not necessarily discrete, a similar more general representation holds, see Spreij (1998).

The *observation* or *output* process Y, a *random* transformation of the state process, takes its values in the set $F = \{f_1, \dots, f_m\}$ of basis vectors of $\mathbb{R}^m$. The matrices $\{H_t\}$ form an iid sequence, independent of $\{X_t\}$, and each column of any of these matrices is assumed to be a random element of F. Their common distribution is specified by the expectation $EH_t = G$. Clearly, each H_t is the incidence matrix of a random map from E into F. Indeed, if $Y_t = h_t(X_t)$, with the h_t random map from E into F, then we can write $Y_t = \sum_{i=1}^n h_t(e_i) 1_{\{X_t = e_i\}}$. So we define $H_t = [h_t(e_1), \dots, h_t(e_n)]$ to get (1.2). We assume (without loss of generality) the non-degeneracy condition that none of the rows of G is zero.

Define the filtration $\mathbb{F} = \{\mathscr{F}_t\}$ by $\mathscr{F}_t = \sigma\{X_0, \dots, X_t, H_0, \dots, H_t\}$. Clearly, both X and Y are adapted to this filtration, and so is the sequence $\{\varepsilon_t\}$ which is even a martingale difference sequence w.r.t $\mathbb{F}$, because of the independence of the sequences $\{X_t\}$ and $\{H_t\}$. The conditional probabilities in the next proposition are essentially as in Baum and Petrie (1966).

Proposition 1.1. *The joint process (X_t, Y_t) is Markov with respect to $\mathbb{F}$ and the conditional transition probabilities are given by*

$$P(X_t = e_i, Y_t = f_j | \mathscr{F}_{t-1}) = e_i^{\mathrm{T}} \operatorname{diag}(AX_{t-1}) G^{\mathrm{T}} f_j. \tag{1.3}$$

Proof. Notice first that the indicator of the event $\{X_t = e_i, Y_t = f_j\}$ equals $e_i^{\mathrm{T}} X_t Y_t^{\mathrm{T}} f_j$. Hence we can rewrite the conditional probability in Eq. (1.3) as $E[e_i^{\mathrm{T}} X_t Y_t^{\mathrm{T}} f_j | \mathscr{F}_{t-1}]$. So we compute

$$\begin{aligned} E[X_t Y_t^{\mathrm{T}} | \mathscr{F}_{t-1}] &= E[X_t X_t^{\mathrm{T}} H_t^{\mathrm{T}} | \mathscr{F}_{t-1}] \\ &= E[E[X_t X_t^{\mathrm{T}} H_t^{\mathrm{T}} | \mathscr{F}_{t-1}, H_t] | \mathscr{F}_{t-1}] \\ &= E[E[X_t X_t^{\mathrm{T}} | \mathscr{F}_{t-1}, H_t] H_t^{\mathrm{T}} | \mathscr{F}_{t-1}] \\ &= E[E[\operatorname{diag}(X_t) | \mathscr{F}_{t-1}, H_t] H_t^{\mathrm{T}} | \mathscr{F}_{t-1}] \\ &= E[\operatorname{diag}(AX_{t-1}) H_t^{\mathrm{T}} | \mathscr{F}_{t-1}] \\ &= \operatorname{diag}(AX_{t-1}) E[H_t^{\mathrm{T}} | \mathscr{F}_{t-1}] \\ &= \operatorname{diag}(AX_{t-1}) G^{\mathrm{T}}. \end{aligned}$$

The result follows. □

We give an alternative expression for the matrix of one step transition probabilities of the joint chain (X, Y). The state space of this chain consists of all the nm pairs (e_i, f_j). These are renamed and ordered as follows: $s_{(j-1)n+i} = (e_i, f_j)$ for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. Clearly, the map $(i, j) \mapsto (i-1)m + j$ is bijective from $\{1, \dots, n\} \times \{1, \dots, m\}$ onto $\{1, \dots, nm\}$.

Instead of working with (X, Y) we will use the chain Z that carries the same information and which is defined by $Z_t = \operatorname{vec}(X_t Y_t^{\mathrm{T}})$. Recall that the vec-operator applied to a matrix results in a vector where all the columns of this matrix are stacked one beneath the other (Magnus and Neudecker, 1988, p. 30). Working with the process Z has the advantage that it has a similar representation as Eq. (1.1), see Eq. (1.11) below.

Clearly, the state space of Z is the set of basis vectors of $\mathbb{R}^{nm}$. If we call this set $\{z_1, \dots, z_{nm}\}$ we see that $(X_t, Y_t) = s_k$ iff $Z_t = z_k$. Notice also the following relations. $Z_t = Y_t \otimes X_t$, $X_t = (\mathbf{1}_m^{\mathrm{T}} \otimes I_n) Z_t$ and $Y_t = (I_m \otimes \mathbf{1}_n^{\mathrm{T}}) Z_t$. Here I_m is the m-dimensional identity matrix and $\mathbf{1}_n$ is the n-dimensional column vector with all its elements equal to one.

According to Proposition 1.1, we now get that the $nm \times nm$ matrix Q of transition probabilities of Z can be decomposed as a matrix with 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 the following notation. Let $\Delta(G)$ be the $nm \times n$ matrix defined by

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

In the next lemma we gather some computational results for the delta-operator. These will be used throughout the rest of the paper.

Lemma 1.2. *For any matrices $G \in \mathbb{R}^{m\times n}$, $M \in \mathbb{R}^{p\times m}$ and $N \in \mathbb{R}^{p\times n}$ and for any vectors $w \in \mathbb{R}^n$, $v \in \mathbb{R}^m$ we have*

$$MG = (M \otimes \mathbf{1}_n^{\mathrm{T}})\Delta(G), \tag{1.4}$$

$$(\mathbf{1}_m^{\mathrm{T}} \otimes N)\,\mathrm{diag}(\mathrm{vec}(G^{\mathrm{T}})) = N\Delta(G)^{\mathrm{T}}, \tag{1.5}$$

$$\mathrm{diag}(w)G^{\mathrm{T}}\,\mathrm{diag}(v) = \Delta(G)^{\mathrm{T}}(\mathrm{diag}(v) \otimes w), \tag{1.6}$$

$$(I_m \otimes \mathrm{diag}(w))\,\mathrm{vec}(G^{\mathrm{T}}) = \Delta(G)w, \tag{1.7}$$

$$N\,\mathrm{diag}(\mathbf{1}_m^{\mathrm{T}}G) = (\mathbf{1}_m^{\mathrm{T}} \otimes N)\Delta(G), \tag{1.8}$$

$$\mathrm{vec}(\mathrm{diag}(w)G^{\mathrm{T}}) = \Delta(G)w. \tag{1.9}$$

Proof. By direct calculation. □

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

$$Q = \Delta(G)A(\mathbf{1}_m^{\mathrm{T}} \otimes I_n). \tag{1.10}$$

This expression for Q can also be obtained as follows. By definition of Q we have $E[Z_{t+1}|\mathscr{F}_t] = QZ_t$. So we compute the conditional expectation

$$\begin{aligned} E[Z_{t+1}|\mathscr{F}_t] &= E[\mathrm{vec}(X_{t+1}Y_{t+1}^{\mathrm{T}})|\mathscr{F}_t] \\ &= \mathrm{vec}(E[X_{t+1}Y_{t+1}^{\mathrm{T}}|\mathscr{F}_t]) \\ &= \mathrm{vec}(E[X_{t+1}X_{t+1}^{\mathrm{T}}H_{t+1}^{\mathrm{T}}|\mathscr{F}_t]) \\ &= \mathrm{vec}(\mathrm{diag}(AX_t)G^{\mathrm{T}}) \\ &= (I_m \otimes \mathrm{diag}(AX_t))\,\mathrm{vec}(G^{\mathrm{T}}) \\ &= \Delta(G)AX_t \\ &= \Delta(G)A(\mathbf{1}_m^{\mathrm{T}} \otimes I_n)Z_t \\ &= QZ_t. \end{aligned}$$

Here we used, in the fifth equality, a known result for the vec-operator of the product of three matrices (see Magnus and Neudecker, 1988, p. 30) and in the sixth equality equation (1.7).

From this computation it follows that Z can be represented by

$$Z_t = QZ_{t-1} + \zeta_t, \tag{1.11}$$

where ζ is the martingale difference sequence w.r.t. $\mathbb{F}$.

Furthermore, if $p_0 = EX_0$ and $P_0 = \mathrm{diag}(p_0)$, then the initial distribution of Z is given by the vector $EZ_0 = \mathrm{vec}(P_0G^{\mathrm{T}})$. Indeed, $EZ_0 = E\,\mathrm{vec}(X_0Y_0^{\mathrm{T}}) = \mathrm{vec}(E\,\mathrm{diag}(X_0)H_0^{\mathrm{T}}) = \mathrm{vec}(P_0G^{\mathrm{T}})$. Notice that $\mathrm{vec}(P_0G^{\mathrm{T}}) = \Delta(G)p_0$.

Similarly, we have that if X has an invariant probability vector π, then $\Delta(G)\pi$ is an invariant probability vector for Z. Indeed,

$$Q\Delta(G)\pi = \Delta(G)A(\mathbf{1}_m^{\mathrm{T}} \otimes I_n)\Delta(G)\pi = \Delta(G)A\,\mathrm{diag}(\mathbf{1}_m^{\mathrm{T}}G)\pi,$$

because of (1.8), but since the column sums of G are all equal to one, this is nothing else but

$$\Delta(G)A\,\mathrm{diag}(\mathbf{1}_n^{\mathrm{T}})\pi = \Delta(G)A\pi = \Delta(G)\pi.$$

We close this section with the observation that the law of the bivariate process (X, Y) (equivalently that of Z) is completely determined by the parameters A, G and p_0. This follows from Eq. (1.11) and the observation that ζ is also a martingale difference sequence with respect to the filtration generated by (X, Y). Hence, any process (X, Y) with this law can be thought of as being generated by the set of equations (1.1) and (1.2).

2. On the Markovian character of Y

In this section, we put forward a set of simple conditions such that the process Y is Markov again. We use (in this section, cf. Section 3 for an alternative approach) the results of Rubino and Sericola (1991) who studied the problem of *weak lumpability* of finite Markov chains. This problem is simply to answer the following question: under what condition is a given (deterministic) transformation of a Markov chain again Markov? It has been known for a long time that the Markovian behaviour of the transformed process in general depends on the initial distribution of the chain. If one requires the function of a Markov chain also to be Markov for any choice of the initial distribution, then the chain is said to be *strongly lumpable*. These problems have been studied in the literature for a long time. Apart from the already mentioned paper (Rubino and Sericola, 1991) there is a rather old paper (Burke and Rosenblatt, 1958) and a discussion of these issues in the books by Kemeny and Snell (1960) and Rosenblatt (1974). Recently, Ledoux (1995) and Ball and Yeo (1993) dealt with these problems for denumerable Markov chains.

In the continuous time case we mention the paper by Rubino and Sericola (1993), where a uniformization approach is used and the results of Rubino and Sericola (1991), and the paper by Rogers and Pitman (1981) for Markov processes with an arbitrary state space. The main condition for a transformed process to be Markov, in this paper (condition (5)) applied to finite state processes, is nothing else but the known criterion (3) on Kemeny and Snell (1960, p. 135).

Rubino and Sericola (1991) characterized the set of all initial distributions for which the deterministically transformed process is Markov, which they call $\mathscr{A}_{\mathscr{M}}$, and showed that this characterization is finite dimensional.

We use their result as follows. We view the process Y (being a *random* transformation of X) as a deterministic transformation of the jointly Markov process (X, Y), equivalently of the process Z as in the first section, and we require that Y is Markov for any initial distribution p_0 of X. Notice that this is not the same as requiring that Y has to be Markov for any choice of the distribution of (X_0, Y_0) or of Z_0, since Y_0 depends, in our context, in a special way on X_0 via Eq. (1.2) for $t = 0$. Specifically, if the initial distribution of X is p_0, so $EX_0 = p_0$, then the initial distribution of Y is given by $EY_0 = E[H_0X_0] = E[H_0]E[X_0] = Gp_0$. The idea then is to give conditions under which the vector $(p_0^{\mathrm{T}}, (Gp_0)^{\mathrm{T}})$ or—with a little abuse of notation—the vector $EZ_0 = \mathrm{vec}(P_0G^{\mathrm{T}})$ belongs to the set $\mathscr{A}_{\mathscr{M}}$ for any p_0.

Since we will use the results of Rubino and Sericola (1991), we recall some facts of this paper, and translate these into the terminology that we use (which also leads to a more transparent notation).

Suppose that one is given a homogeneous Markov chain ζ with state space $\{1,\dots,N\}$ and matrix of transition probabilities Q. Let b be a map defined on the state space onto $\{1,\dots,M\}$, where $M \leqslant N$. The problem under consideration is whether the process $\eta = b(\zeta)$ is again a Markov chain [in Rubino and Sericola (1991) this process is called the lumped chain]. Introduce the incidence matrix B with elements $B_{ij} = 1_{\{b(j)=i\}}$. Let $B(i)$ be the ith row of B. Let $p \in \mathbb{R}^N$ be a probability vector, the set of all such probability vectors is denoted by $\mathscr{A}$. By $\mathscr{A}_+$ we denote the subset of $\mathscr{A}$ consisting of the probability vectors that have strictly positive entries only. By $p^{B(i)}$ we denote the probability vector $(B(i)p)^{-1}\operatorname{diag}(B(i))p$ for $i=1,\dots,M$. Assume that ζ admits an invariant distribution whose probability vector is denoted by π.

For any vector $p \in \mathscr{A}$, we write $P = \operatorname{diag}(p)$ and we denote by B_p^+ the right pseudo-inverse of B defined by $B_p^+ = PB^{\mathrm{T}}(BPB^{\mathrm{T}})^{-1}$ if BPB^{T} is invertible. In particular, if the chain is irreducible then π is unique and all its elements are strictly positive, so that B_π^+ is well defined.

In Rubino and Sericola (1991) the following sets are defined (although described in a rather different way): $\mathscr{A}^1 = \{p \in \mathscr{A}\colon BQB_p^+ = BQB_\pi^+\}$, and recursively for $j \geqslant 2$, $\mathscr{A}^j = \{p \in \mathscr{A}^{j-1}\colon Qp^{B(i)} \in \mathscr{A}^{j-1}\ \forall i = 1,\dots,M\}$. In Rubino and Sericola (1991) the following result is proved.

Lemma 2.1. *If p corresponds to the initial distribution of ζ, then η is a Markov chain iff $p \in \bigcap_{j=1}^N \mathscr{A}^j = \mathscr{A}^N$. So $\mathscr{A}_{\mathscr{M}} = \mathscr{A}^N$.*

In the same paper one can find also an algorithm that determines this intersection in at most N steps.

In the next theorem we formulate a necessary and sufficient condition for Y to be Markov. Later on we will use Theorem 3.1 for an alternative proof. Before stating Theorem 2.2 we want to stress that, although in principle Lemma 2.1 contains the complete solution of the problem, one should find necessary and sufficient conditions in terms of the parameters A and G of the bivariate process (X,Y) and this is precisely what Theorem 2.2 tells us.

Theorem 2.2. *The process Y is Markov for all initial distributions p_0 of X iff the following condition holds: the matrix*

$$GAPG^{\mathrm{T}}\operatorname{diag}(Gp)^{-1} \tag{2.1}$$

with $P = \operatorname{diag}(p)$ is independent of the vector $p \in \mathscr{A}_+$.

Equivalently, iff the map $M : \mathscr{A}_+ \to \mathbb{R}^{m\times m}$ defined by $M(p) = GAPG^{\mathrm{T}}\operatorname{diag}(Gp)^{-1}$ is constant. If Y is Markov, then the matrix of one step transition probabilities is given by the common value of (2.1) for $p \in \mathscr{A}_+$.

Proof. The proof is an application of Lemma 2.1. So we take ζ equal to Z, $Q = \Delta(G)A(\mathbf{1}_m^{\mathrm{T}} \otimes I_n)$, $N = nm$, $M = m$, $B = I_m \otimes \mathbf{1}_n^{\mathrm{T}}$. The first thing to do is to find an expression for BQB_p^+ for p in the nm-dimensional simplex and for the special form $p = \Delta(G)p_0$ where p_0 is the initial probability vector of X, since *all initial distributions of Z are of this type* (see Section 1).

All the matrix computations that are involved here are rather straightforward, but tedious. Therefore, we only give some intermediate results, which are the following. For $p = \Delta(G)p_0$ and $P = \operatorname{diag}(p)$ we get $BPB^{\mathrm{T}} = \operatorname{diag}(Gp_0)$.

Next, we compute $BQPB^{\mathrm{T}} = (I_m \otimes \mathbf{1}_n^{\mathrm{T}})\Delta(G)A(\mathbf{1}_m^{\mathrm{T}} \otimes I_n)P(I_m \otimes \mathbf{1}_n)$. In order to get a decent expression for this we use some properties concerning the Δ-operator.

Applying (1.4) with $M = I_m$ and (1.5) with $N = P_0$ $(P = (I_m \otimes P_0)\operatorname{diag}(\operatorname{vec}(G^{\mathrm{T}})))$, we get

$$\begin{aligned} BQPB^{\mathrm{T}} &= (I_m \otimes \mathbf{1}_n^{\mathrm{T}})\Delta(G)A(\mathbf{1}_m^{\mathrm{T}} \otimes I_n)P(I_m \otimes \mathbf{1}_n) \\ &= GA(\mathbf{1}_m^{\mathrm{T}} \otimes I_n)(I_m \otimes P_0)\operatorname{diag}(\operatorname{vec}(G^{\mathrm{T}}))(I_m \otimes \mathbf{1}_n) \end{aligned}$$

$$
\begin{aligned}
&= GA(\mathbf{1}_m^{\mathrm{T}} \otimes P_0)\,\mathrm{diag}(\mathrm{vec}(G^{\mathrm{T}}))(I_m \otimes \mathbf{1}_n)\\
&= GAP_0\Delta(G)^{\mathrm{T}}(I_m \otimes \mathbf{1}_n)\\
&= GAP_0G^{\mathrm{T}}.
\end{aligned}
$$

So, we get

$$BQB_p^+ = GAP_0G^{\mathrm{T}}\,\mathrm{diag}(Gp_0)^{-1}. \tag{2.2}$$

Next, we are going to check whether $Qp^{B(i)} \in \mathscr{A}^1$. First we determine $p^{B(i)}$. Since $B(i) = f_i^{\mathrm{T}} \otimes \mathbf{1}_n^{\mathrm{T}}$, we get $p^{B(i)} = (G_{i.}p_0)^{-1}f_i \otimes P_0G_{i.}^{\mathrm{T}}$.

It is convenient to introduce the following notation. Let $d_i(p_0)$=$AP_0G_{i.}^{\mathrm{T}}(G_{i.}p_0)^{-1}$, and $D_i(p_0)$=diag$(d_i(p_0))$. Then

$$
\begin{aligned}
Qp^{B(i)} &= \Delta(G)A(\mathbf{1}_n^{\mathrm{T}} \otimes I_m)(G_{i.}p_0)^{-1}f_i \otimes P_0G_{i.}^{\mathrm{T}}\\
&= \Delta(G)d_i(p_0)\\
&= (I_m \otimes D_i(p_0))\,\mathrm{vec}(G^{\mathrm{T}}).
\end{aligned} \tag{2.3}
$$

Then, it easily follows that $B\,\mathrm{diag}(Qp^{B(i)})B^{\mathrm{T}} = \mathrm{diag}(Gd_i(p_0)) = \mathrm{diag}(GAP_0G_{i.}^{\mathrm{T}})(G_{i.}p_0)^{-1}$, and $BQ\,\mathrm{diag}(Qp^{B(i)})$ $B^{\mathrm{T}} = GAD_i(p_0)G^{\mathrm{T}}$. Hence,

$$BQB_{Qp^{B(i)}}^+ = GAD_i(p_0)G^{\mathrm{T}}(\mathrm{diag}(Gd_i(p_0)))^{-1} \tag{2.4}$$

$$=GA\,\mathrm{diag}(AP_0G_{i.}^{\mathrm{T}})G^{\mathrm{T}}(\mathrm{diag}(GAP_0G_{i.}^{\mathrm{T}}). \tag{2.5}$$

Now we make the following observation. For each i and all p_0, the vector $d_i(p_0)$ is again a probability vector. Hence, since by assumption the map M is constant, the resulting matrix in (2.4) is the same as in (2.2). It follows that $p = \Delta(G) \in \mathscr{A}^1$ implies $Qp^{B(i)} \in \mathscr{A}^1$, or in other words $p \in \mathscr{A}^2$. But then, obviously, all the $\mathscr{A}^j$ are the same, which concludes the proof. □

The condition that (2.1) shares the same value for all p can be verified by means of the next corollary. Recall that the Hadamard product $u \odot v$ of two vectors u and v of the same dimension is defined as the vector with the product of the ith components of u and v as the ith component.

Corollary 2.3. *The process Y is Markov for any initial distribution of X iff the ith and jth columns of the matrix GA are equal, whenever $G_{.i} \odot G_{.j} \neq 0$.*

Proof. The condition that (2.1) is constant in p is equivalent to the requirement that $GA[\mathrm{diag}(p)G^{\mathrm{T}}\,\mathrm{diag}(Gq) - \mathrm{diag}(q)\,\mathrm{diag}(Gp)] = 0$ for any $p, q \in \mathscr{A}_+$. The expression on the left-hand side of this equation is bilinear in p and q and hence determined by taking for p and q basis vectors e_i and e_j of $\mathbb{R}^n$, in which case we get, with $G_{.i}$ and $G_{.j}$ the ith and jth columns of G respectively,

$$GA[e_iG_{.i}^{\mathrm{T}}\,\mathrm{diag}(G_{.j}) - e_jG_{.j}^{\mathrm{T}}\,\mathrm{diag}(G_{.i})] = 0. \tag{2.6}$$

Because $G_{.i}^{\mathrm{T}}\,\mathrm{diag}(G_{.j}) = G_{.j}^{\mathrm{T}}\,\mathrm{diag}(G_{.i}) = (G_{.i} \odot G_{.j})^{\mathrm{T}}$, Eq. (2.6) becomes

$$GA(e_i - e_j)(G_{.i} \odot G_{.j})^{\mathrm{T}} = 0, \tag{2.7}$$

and hence we obtain $GA(e_i - e_j) = 0$ if $G_{.i} \odot G_{.j} \neq 0$, which proves the corollary. □

Example. Let

$$A=\begin{bmatrix}\frac{1}{2} & \frac{1}{3} & \frac{1}{6}\\ \frac{1}{4} & \frac{1}{3} & \frac{5}{12}\\ \frac{1}{4} & \frac{1}{3} & \frac{5}{12}\end{bmatrix} \quad\text{and}\quad G=\begin{bmatrix}\frac{1}{2} & \frac{2}{3} & \frac{1}{3}\\ \frac{1}{2} & \frac{1}{3} & \frac{2}{3}\end{bmatrix}.$$

Then

$$GA=\begin{bmatrix}\frac{1}{2} & \frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2}\end{bmatrix}.$$

Hence Y is a Markov chain for all initial conditions of X. Its transition matrix is

$$\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{bmatrix}.$$

It is easy to check that no deterministic function of X is a Markov chain.

Like in Rubino and Sericola (1991) one can also ask the question 'for which initial distributions of X is the process Y Markov?' The answer is given below. We omit the proof, but return to this in Section 3. We need some additional notation.

Let $\mathscr{I}$ be an arbitrary point in $\{1,\ldots,m\}^n$, $\mathscr{I}=(i_1,\ldots,i_n)$ say. Define the maps $d_{\mathscr{I},k}$: $\mathscr{A}\to\mathscr{A}$ by $d_{\mathscr{I},0}(p)=p$ and for $k\geqslant 1$ recursively by

$$d_{\mathscr{I},k}(p)=A\,\mathrm{diag}(d_{\mathscr{I},k-1}(p))G_{i_k\cdot}^{\mathrm{T}}(G_{i_k\cdot}d_{\mathscr{I},k-1}(p))^{-1}.$$

Theorem 2.4. *The process Y is Markov if the initial distribution p_0 of X satisfies for all $\mathscr{I}\in\{1,\ldots,m\}^n$ and for all $k=0,\ldots,n$ the following condition*:

$$GA\,\mathrm{diag}(d_{\mathscr{I},k}(p_0))G^{\mathrm{T}}\,\mathrm{diag}(Gd_{\mathscr{I},k}(p_0))^{-1}=GA\,\mathrm{diag}(\pi)G^{\mathrm{T}}\,\mathrm{diag}(G\pi)^{-1}. \tag{2.8}$$

Remark. The condition of this theorem is equivalent to requiring that each of the vectors $d_{\mathscr{I},k}(p_0)$ belongs to the kernel of the matrix

$$(I_m\otimes GA)\,\mathrm{diag}(\mathrm{vec}\,G^{\mathrm{T}})(G\pi\otimes I_n-G\otimes\pi). \tag{2.9}$$

As in Rubino and Sericola (1991) this requirement can be transformed into a system of linear equations that p_0 is supposed to satisfy. Hence the set of p_0 satisfying this system of equations is the intersection (possibly empty) of a linear subspace of $\mathbb{R}^n$ and the n-dimensional simplex.

3. Filtering

In this section we give some filtering and prediction formulas. By the filtering problem for a hidden Markov chain (X,Y) we mean the determination for each t of the conditional law of X_t given $Y_0,\ldots,Y_t$. Apart from the fact that these formulas are of interest in their own right, they are also helpful in studying the Markovian character of a hidden Markov chain. This has been observed implicitly by Rogers and Pitman (1981), where they write (p. 574) that their criterion (1) for $Y=\phi\circ X$ (here ϕ is a deterministic transformation) to be Markov is unsatisfactory, since '*one has to be able to calculate the conditional distribution of X_t given the whole history of $\phi\circ X$ up to time t*', but this is just rather easy in our case where the state space is finite.

For each t we denote by $\mathscr{F}_t^Y$, the σ-algebra generated by $Y_0,\ldots,Y_t$. Since the state space of X is a set of basis vectors, the conditional law of X_t given $Y_0,\ldots,Y_t$, is completely determined by the conditional expectation

$E[X_t|\mathscr{F}_t^Y]$. 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 $E[X_{t+1}|\mathscr{F}_t^Y]$. We will use the notations $E[X_t|\mathscr{F}_t^Y]=\hat{X}_t$ and $E[X_{t+1}|\mathscr{F}_t^Y]=\hat{X}_{t+1|t}$. Similarly we write $E[Y_{t+1}|\mathscr{F}_t^Y]=\hat{Y}_{t+1|t}$.

In the book by Elliott et al. (1995), recursive formulae for unnormalized filters are obtained by a measure transformation. It is possible to obtain in particular Eq. (3.2) below from equation (4.3) on p. 28 of Elliott et al. (1995). For the convenience of the reader we give in the appendix an elementary proof based on an approach, that directly leads to a simple recursive formula for the (normalized) conditional probabilities itself, instead of a recursion for the unnormalized conditional probabilities.

We return to the setting of Section 1 and we give the recursive filtering formula for the stochastic system with the HMC Y as its output. The following holds.

Theorem 3.1. (i) *The conditional distribution of the X_t given $Y_0,\dots,Y_t$ is recursively determined by*

$$\hat{X}_t = \operatorname{diag}(A\hat{X}_{t-1})G^{\mathrm{T}}\operatorname{diag}(GA\hat{X}_{t-1})^{-1}Y_t, \tag{3.1}$$

with initial condition $\hat{X}_0 = P_0G^{\mathrm{T}}\operatorname{diag}(Gp_0)^{-1}Y_0$. Here $p_0 = EX_0$ and $P_0 = \operatorname{diag}(p_0)$.

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

$$\hat{X}_{t+1|t} = A\operatorname{diag}(\hat{X}_{t|t-1})G^{\mathrm{T}}\operatorname{diag}(G\hat{X}_{t|t-1})^{-1}Y_t, \tag{3.2}$$

with initial condition $X_{0|-1} = EX_0 = p_0$.

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

$$\hat{Y}_{t+1|t} = GA\operatorname{diag}(\hat{X}_{t|t-1})G^{\mathrm{T}}\operatorname{diag}(G\hat{X}_{t|t-1})^{-1}Y_t. \tag{3.3}$$

Remark. If we define for $x \in \mathbb{R}_+^n$ the matrix $G_x := \operatorname{diag}(x)G^{\mathrm{T}}\operatorname{diag}(Gx)^{-1}$, then Eqs. (3.1)–(3.3) 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 if 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.

We return to the questions posed in Section 2. The condition that (2.1) is constant in p in Theorem 2.2 was seen to be necessary and sufficient for Y to be Markov for any initial condition of X. In view of Eq. (3.3) this is no surprise. We see that the conditional expectation $Y_{t+1|t}$ depends on $Y_0,\dots,Y_{t-1}$ through the matrix $GAG_{\hat{X}_{t|t-1}}$, using previously introduced notation. Hence we have the Markov property for Y if and only if this matrix is independent of the specific values of $\hat{X}_{t|t-1}$. In particular, if we allow the initial condition $X_{0|-1} = EX_0$ to be arbitrary, then we have that Y is Markov iff the matrix GAG_p is independent of p, which is again the condition of Theorem 2.2.

Using the expression (3.3) it is now also possible to give an answer (although less elegant) to the other question posed in Section 2, namely which initial distributions of X yield the process Y Markov?

It was observed in Kemeny and Snell (1960) that if there exists an initial probability vector p that yields a deterministic function of X Markov, then the same is true for the initial vector Ap and hence for an invariant vector π, if there exists one. We claim that this is also true in our case, where Y is a random function of X, because of the following argument. Let p be an initial probability vector of X, then $Z = Y \otimes X$ has initial probability vector $\Delta(G)p$ (cf. Section 1). We have also seen that $Q\Delta(G)p = \Delta(G)Ap$, which is the initial probability vector of Z if X has Ap as initial distribution. Since Y is a deterministic function of Z, we can apply the observation in Kemeny and Snell (1960) to get our claim.

For a compact condition on p we introduce some auxiliary notation. Define for $p \in \mathscr{A}_+$ the matrix $T(p) = AG_p = A\operatorname{diag}(p)G^{\mathrm{T}}\operatorname{diag}(Gp)^{-1} \in \mathbb{R}^{n\times m}$. The prediction equation (3.2) tells us that if we start the recursion in a vector p, then $\hat{X}_{1|0} = T(p)Y_0$ and from (3.3) we get $\hat{Y}_{1|0} = GT(p)Y_0$. Consider now the columns

$T(p)_i$ $(i=1,\dots,m)$ of $T(p)$. If $Y_0=f_i$ is realized, then we get $\hat{X}_{2|1}=T(T(p)_i)Y_1$ and $\hat{Y}_{2|1}=GT(T(p)_i)Y_1$. So if we want that Y is Markov we need to have that for all i the matrix $GT(T(p)_i)$ is equal to the matrix $GT(p)$. This can be continued for the next incoming observations Y_t, $t\geqslant 2$, and the resulting condition on p that yields Y a Markov chain is the one described in Theorem 2.4.

A relatively simple sufficient condition on p that yields Y Markov is

$$T(T(p)_i)=T(p)\quad \text{for all } i=1,\dots,m. \tag{3.4}$$

Indeed if (3.4) holds, then we have $\hat{X}_{t+1|t}=T(p)Y_t$ for all t, and hence $\hat{Y}_{t+1|t}=GT(p)Y_t$. Notice that it is possible to rewrite (3.4) explicitly as a set of linear equations in p.

Appendix A. Proof of the filter formulas

For $\hat{X}_t=E[X_t|\mathscr{F}_t^Y]$ we also write $E[X_t|Y^t]$. Here Y^t is a vector that represents all the observations $Y_0,\dots,Y_t$. One possible definition—which is the one we take—of Y^t is a recursive one: $Y^0=Y_0$ and for $t\geqslant 1$ we have $Y^t=Y_t\otimes Y^{t-1}$. So Y^t is m^{t+1}-dimensional. First we use the well-known result for conditional expectations given a finitely generated σ-field or a random variable with a finite range. With the presently introduced notation, this result can be conveniently formulated for our problem as

$$\hat{X}_t=E[X_t(Y^t)^{\mathrm{T}}]\,\mathrm{diag}(EY^t)^{-1}Y^t. \tag{A.1}$$

For simplicity we assume that the inverse in (A.1) exists, although similar calculations can be carried out by replacing it with the Moore–Penrose inverse, or equivalently by deleting the entries of the vector Y^t that have zero expectation as well as the corresponding entries of EY^t.

Let $K_t=E[X_t(Y^t)^{\mathrm{T}}]$. We give first a recursive formula for K_t.

$$K_t=\Delta(G)^{\mathrm{T}}(I_m\otimes AK_{t-1}). \tag{A.2}$$

Furthermore, $EY^t=E[Y^tX_t^{\mathrm{T}}]\mathbf{1}_n=K_t^{\mathrm{T}}\mathbf{1}_n=(I_m\otimes(AK_{t-1})^{\mathrm{T}})\,\mathrm{vec}(G^{\mathrm{T}})$, which equals $\mathrm{vec}(GAK_{t-1})$ by a familiar rule for the vec-operator (see Magnus and Neudecker, 1988, p. 31). Write now $D_t=\mathrm{diag}(EY_t)$. One can check that

$$D_t^{-1}Y^t=(I_m\otimes Y^{t-1})\,\mathrm{diag}(GAK_{t-1}Y^{t-1})^{-1}Y_t. \tag{A.3}$$

Then we use Eqs. (A.1) and (A.3) to write

$$\begin{aligned}\hat{X}_t&=\Delta(G)^{\mathrm{T}}(I_m\otimes AK_{t-1})(I_m\otimes Y^{t-1})\,\mathrm{diag}(GAK_{t-1}Y^{t-1})^{-1}Y_t\\&=\Delta(G)^{\mathrm{T}}(I_m\otimes AK_{t-1}Y^{t-1})\,\mathrm{diag}(GAK_{t-1}Y^{t-1})^{-1}Y_t\\&=\mathrm{diag}(AK_{t-1}Y^{t-1})G^{\mathrm{T}}\,\mathrm{diag}(GAK_{t-1}Y^{t-1})^{-1}Y_t.\end{aligned} \tag{A.4}$$

Eq. (A.4) follows from application of (1.6) with $v=\mathbf{1}_m$ and the preceding equality follows from the multiplication rule for Kronecker products (Magnus and Neudecker, 1988, p. 28). Now we make the following observations. For any non-zero number d, it holds that $\mathrm{diag}(w)G^{\mathrm{T}}\,\mathrm{diag}(v)^{-1}=\mathrm{diag}(dw)G^{\mathrm{T}}\,\mathrm{diag}(dv)^{-1}$. We apply this observation to equation (A.4) with $d=(Y^{t-1})^{\mathrm{T}}D_{t-1}^{-1}Y^{t-1}$ to obtain

$$\hat{X}_t=\mathrm{diag}(AK_{t-1}D_{t-1}^{-1}Y^{t-1})G^{\mathrm{T}}\,\mathrm{diag}(GAK_{t-1}D_{t-1}^{-1}Y^{t-1})^{-1}Y_t. \tag{A.5}$$

But this is exactly Eq. (3.1) in view of Eq. (A.1) with $t-1$ instead of t. □

References

Ball, F., Yeo, G.F., 1993. Lumpability and marginalisability for continuous-time Markov chains. J. Appl. Probab. 19, 518–528.

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

Burke, C.J., Rosenblatt, M., 1958. A Markovian function of a Markov chain. Ann. Math. Statist. 29, 1112–1122.

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

Kemeny, J.G., Snell, J.L., 1960. Finite Markov Chains. Van Nostrand, Princeton, NJ.

Ledoux, J., 1995. On weak lumpability of denumerable Markov chains. Statist. Probab. Lett. 25, 329–339.

Magnus, J.R., Neudecker, H., 1988. Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, New York.

Rogers, L.C.G., Pitman, J.W., 1981. Markov functions. Ann. Probab. 9 (4), 573–582.

Rosenblatt, M., 1974. Random Processes, GTM 117. Springer, Berlin.

Rubino, G., Sericola, B., 1991. A finite characterization of weak lumpable Markov processes. Part I: the discrete time case. Stochastic. Proc. Appl. 38, 195–204.

Rubino, G., Sericola, B., 1993. A finite characterization of weak lumpable Markov processes. Part II: the continuous time case. Stochastic. Proc. Appl. 45, 115–125.

Spreij, P.J.C., 1998. A representation result for finite Markov chains. Statist. Probab. Lett. 38, 183–186.