High-Order Extensions of the Double Chain Markov Model

André Berchtold*

Technical Report no. 356
Department of Statistics
University of Washington
Seattle, WA 98195-4322

July 1999

Abstract

The Double Chain Markov Model is a fully Markovian model for the representation of time-series in random environment. In this article, we show that it can handle transitions of high-order between both a set of obsevations and a set of hidden states. In order to reduce the number of parameters, each transition matrix can be replaced by a Mixture Transition Model. We provide a complete derivation of the algorithms needed to compute the model. Three applications, the analysis of a sequence of DNA, the song of the wood pewee and the behavior of young monkeys, show that this model is of great interest for the representation of data which can be decomposed into a finite set of patterns.

Keywords: Double Chain Markov Model (DCMM), High-Order Transitions, Mixture Transition Distribution Model (MTD), Forward-Backward Algorithm, Baum-Welsh Algorithm, Viterbi Algorithm, DNA, Pewee, Behavior.

*Email: berchtol@stat.washington.edu, Andre.Berchtold@themes.unige.ch

Contents

List of Figures

List of Tables

1 Introduction

Several Markovian models can be used for the analysis of time-series in discrete time. Among them, we can cite the standard Markov chain in which the value taken by a random variable X at time t is explained by the value taken by this same variable at time $t-1$ (first-order model) or at time $t-\ell, \ldots, t-1$ (high-order model). This model is completely visible since each state of the process is exactly identified with one value taken by the random variable. It is used in a lot of fields including meteorology, DNA analysis and mobility. See e.g. Dynkin (1965), Kemeny & Snell (1976) and Kijima (1997) for more details.

The Hidden Markov Model (HMM) considers a slightly different process. Instead of observing the variable X governed by the Markov chain, we observe a second random variable Y. To each state of the process (each value of X) corresponds a different probability distribution of the values taken by Y, the distribution effectively used at time t being the one corresponding to the state taken by X_t. In this process, the successive observations of the variable Y are supposed conditionally independent. Speech recognition is the first field of application of HMMs. Rabiner (1989) and MacDonald & Zucchini (1997) are good references.

One of the limitations of HMMs is the assumption of conditional independence of the observations. Another class of models used more particularly in biology remove this constraint. Markov models in random environment consider an observed variable Y whose transition process is represented by several transition matrices. At each time t, a matrix is selected through a decision rule and is used to compute the next observation. See e.g. Cogburn (1984) and Collins & McNamara (1998) for additional information.

Both HMMs and models in random environment have a Markovian part and a non-Markovian one. In Berchtold (1999a), we presented a fully Markovian model called the Double Chain Markov Model (DCMM). It can be viewed either as a HMM with a direct relation between successive observations or as a model under environment in which the decision rule is Markovian. More precisely, the DCMM combines two Markov chains, hence its name: an observed non-homogeneous Markov chain and a hidden homogeneous one whose state at each time t decides of the matrix used in the visible process. Figure 1 presents the basic DCMM with first-order hidden and visible processes.

The idea behind the DCMM is not completely new. Paliwal (1993) proposed a discrete HMM with a direct relation between observations. Wellekens (1987) also presented a similar model for the continuous case. Nevertheless, the presentation given in Berchtold (1999a) is a lot more detailed and shows that this type of model can be used successfully in different fields.

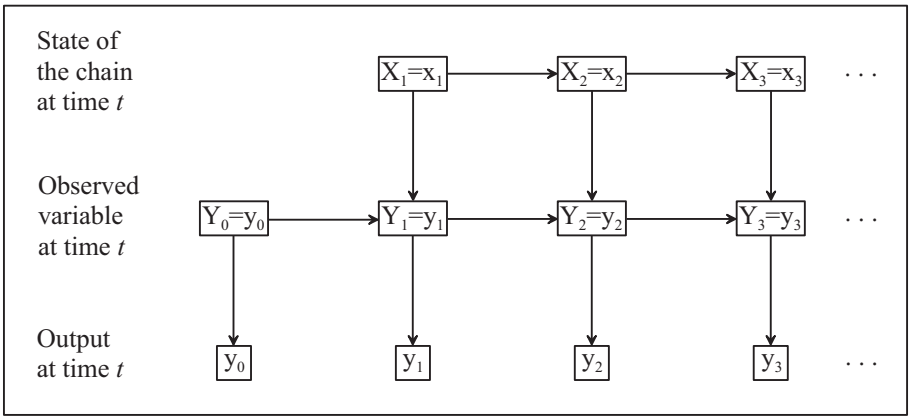

Figure 1. A Double Chain Markov Model with first-order dependences. X denotes the hidden variable and Y the observed process. At time t, X_t depends on X_{t-1} and Y_t depends on both X_t and Y_{t-1}.

In this article, we extend the principle of the DCMM by the use of Markovian dependences of order greater than 1. Let ℓ denotes the order of the dependence between the Xs, and f the order of the dependence between the Ys. Then, X_t depends on $X_{t-\ell}, \ldots, X_{t-1}$ and Y_t depends on X_t and $Y_{t-f}, \ldots, Y_{t-1}$. Figure 2 shows these dependences.

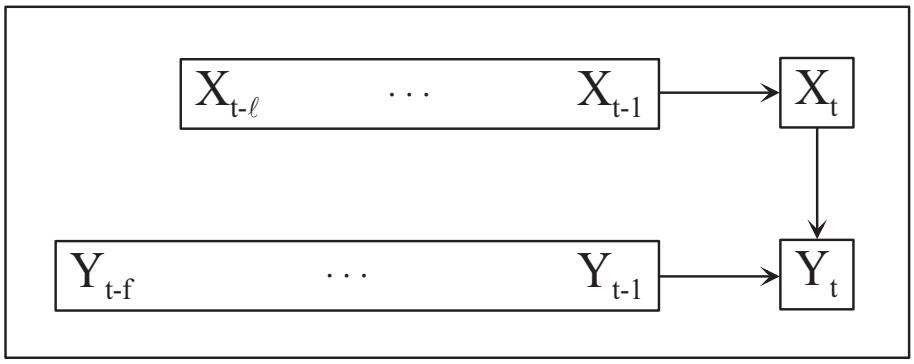

Figure 2. A Double Chain Markov Model with an order ℓ hidden chain and an order f visible chain.

The development of the DCMM was driven by two considerations. Firstly, a lot of datasets present non-homogeneous Markovian dependences, but the number of tools available for their analysis is small. Secondly, rather than considering Markov chains, HMMs and models in random environment as alternative approaches to a same problem, it would be more easy to consider a general model of which they would be particular cases. Then, they could be estimated using a unique set of algorithms and it would become more easy to find the best modeling for a particular set of data. In regard of these considerations, the DCMM is a good answer.

The model is formally defined in Section 2 and its estimation is presented in Section 3. In Section 4, we introduce a modeling principle leading to a substantially more parsimonious model and Section 5 presents several applications of the DCMM. Finally, we provide in appendix the complete derivation of the algorithms of Section 3.

2 Model

We consider a random variable Y taking value in the finite set $\{1, \ldots, K\}$ and a sequence of observations of Y. We make the assumption that the probability to observe a particular value y_t at time t depends on the value of Y at time $t-f$ to $t-1$. Moreover, we suppose that the time-series is non-homogeneous and that the whole sequence is best represented using alternatively M transition matrices of order f. Since Y depends on its past, we need f observations to initialize the process. For convenience purpose, we note $Y_{-f+1}, \ldots Y_0$ these first observations (see Figure 3), and $Y_1, \ldots, Y_T$ the observations used in the computation of the log-likelihood.

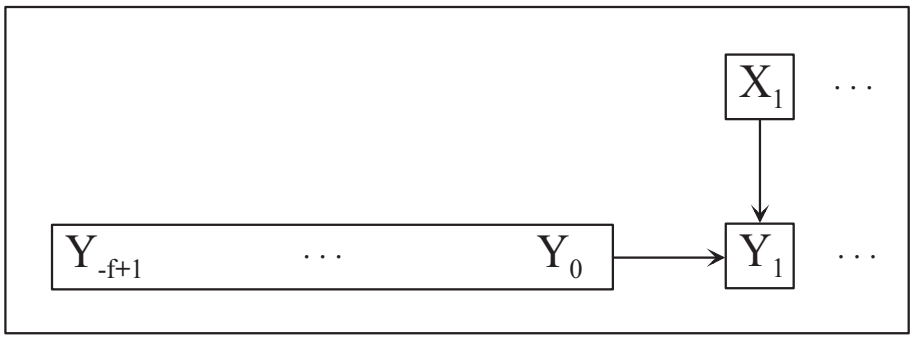

Figure 3. Initialization of the Double Chain Markov Model. To estimate an order f visible chain, we introduce f first observations $Y_{-f+1}, \ldots, Y_0$.

The probability of observing y_t depends not only on the past of the variable Y_t, but also on another unobservable variable X_t taking value in the finite set $\{1, \ldots, M\}$. As for Y_t, the probability to observe x_t depends on the value of $X_{t-\ell}, \ldots, X_{t-1}$, where ℓ is the order of the hidden dependence. Normally, we would need ℓ successive values of X_t to initialize the process, but since this variable is unobserved, we replace these elements by probability distributions. We note π_1 the probability distribution of X_1, $\pi_{2|1}$ the conditional probability distribution of X_2 given X_1, ..., and $\pi_{\ell|1,\ldots,\ell-1}$ the conditional probability distribution of X_ℓ given $X_1, \ldots, X_{\ell-1}$.

The Double Chain Markov Model is completely defined by supposing that the transition matrix used to represent Y_t given its past is chosen in function of the state of X_t. In summary, a DCMM of order ℓ for the hidden chain and f for the visible one is fully described by the following elements:

- A set of hidden states, $\mathcal{S}(X) = \{1, \ldots, M\}$.
- A set of possible outputs, $\mathcal{S}(Y) = \{1, \ldots, K\}$.
- The probability distribution of the first ℓ hidden states given the previous states, π_1, $\pi_{2|1}, \ldots, \pi_{\ell|1,\ldots,\ell-1}$.
- An order ℓ transition matrix between hidden states, $A = \{a_{j_\ell,\ldots,j_0}\}$, $j_\ell, \ldots, j_0 \in \mathcal{S}(X)$.
- A set of order f transition matrices between successive observations of the variable Y given a particular state of X, $C = \{c_{i_f,\ldots,i_0,j_0}\}$, $i_f, \ldots, i_0 \in \mathcal{S}(Y)$, $j_0 \in \mathcal{S}(X)$. C can also be rewritten in a more convenient way as $C = \{C^{(j_0)}\}$, with $C^{(j_0)} = \{c^{(j_0)}_{i_f,\ldots,i_0}\}$.

A DCMM μ is defined as $\mu = \{\pi, A, C\}$. It has $\sum_{g=0}^{\ell-1} M^g(M-1)$ independent parameters for the set of distributions π, $M^\ell(M-1)$ independent parameters for the transition matrix between hidden states A, and $MK^{f-1}(K-1)$ independent parameters for the transition matrices between observations. The total number of parameters can become very large when ℓ or f is greater than 1. It is then useful to replace π, A and C by modelings. We consider this question in Section 4.

3 Estimation

Since the DCMM can be viewed as a generalization of the HMM, similar estimation problems occur and they can be solved in the same way. We consider three different questions:

1. The estimation of the likelihood of the data given a model μ.
2. The estimation of π, A and C given the data.
3. The estimation of the optimal sequence of hidden states given a model μ and the data.

The first problem is solved using a dynamic programming method. The estimation of the parameters is achieved through an Expectation-Maximization (EM) algorithm, and the optimal sequence of states is computed by the Viterbi algorithm. Sections 3.1 to 3.3 provide the main equations of each algorithm, and Section 3.4 presents the case of simultaneous data. The complete derivation of these algorithms is given in appendix.

3.1 Likelihood of the sequence of data

We want to compute the likelihood of the sequence of observations given the model μ:

$$L = P(Y_{-f+1}, \ldots, Y_T|\mu)$$

For lisibility purpose, we write Y_t for $Y_t = y_t$. Moreover, we will not further indicate that the computation is intended given the model μ. This problem can be solved through the "Forward procedure" developped by Rabiner (1989) for the estimation of the HMM. Let

$$\alpha_t(j_{\ell-1}, \ldots, j_0) = P(Y_{-f+1}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0) \tag{1}$$

For $t = 1$ equation (1) becomes

$$\alpha_1(j_0) = c^{(j_0)}_{y_{-f+1},\ldots,y_1}\, \pi_{j_0} \tag{2}$$

For $t = 2$, we have

$$\alpha_2(j_1, j_0) = c^{(j_0)}_{y_{-f+2},\ldots,y_2}\, \pi_{j_0|j_1}\, \alpha_1(j_1) \tag{3}$$

and for $t = 3, \ldots, \ell$, we obtain

$$\alpha_t(j_{t-1}, \ldots, j_0) = c^{(j_0)}_{y_{t-f},\ldots,y_t}\, \pi_{j_0|j_{t-1}\ldots j_1}\, \alpha_{t-1}(j_{t-1}, \ldots, j_1) \tag{4}$$

Finally, for $t = \ell + 1, \ldots, T$,

$$\alpha_t(j_{\ell-1}, \ldots, j_0) = c^{(j_0)}_{y_{t-f},\ldots,y_t} \sum_{j_\ell=1}^{M} a_{j_\ell,\ldots,j_0} \alpha_{t-1}(j_\ell, \ldots, j_1) \tag{5}$$

The likelihood of the entire sequence of observations is obtained by summing $\alpha_T(j_{\ell-1}, \ldots, j_0)$ over $j_{\ell-1}, \ldots, j_0$:

$$L = \sum_{j_{\ell-1},\ldots,j_0=1}^{M} \alpha_T(j_{\ell-1}, \ldots, j_0) \tag{6}$$

The iterative computation of α_t is sufficient to obtain the likelihood. However, we define here another iterative algorithm similar to the "Backward procedure" appearing in Rabiner (1989). It will be used later for the estimation of π, A and C. Let

$$\beta_t(j_{\ell-1}, \ldots, j_0) = P(Y_{t+1}, \ldots, Y_T|Y_{t-f}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0)$$

This definition implies that

$$\beta_T(j_{\ell-1},\ldots,j_0) = 1 \quad , \quad \forall\, j_{\ell-1},\ldots,j_0$$

For $t = T-1,\ldots,\ell$, we have

$$\beta_t(j_{\ell-1},\ldots,j_0) = \sum_{j=1}^{M} a_{j_{\ell-1},\ldots,j_0,j}\, c^{(j)}_{y_{t-f+1},\ldots,y_{t+1}}\, \beta_{t+1}(j_{l-2},\ldots,j_0,j) \tag{7}$$

and for $t = \ell-1,\ldots,1$, we obtain

$$\beta_t(j_{t-1},\ldots,j_0) = \sum_{j=1}^{M} \pi_{j|j_{t-1},\ldots,j_0}\, c^{(j)}_{y_{t-f+1},\ldots,y_{t+1}}\, \beta_{t+1}(j_{t-1},\ldots,j_0,j) \tag{8}$$

With this result, the likelihood can be rewritten as

$$L = \sum_{j_{\ell-1},\ldots,j_0=1}^{M} \alpha_t(j_{l-1},\ldots,j_0)\, \beta_t(j_{l-1},\ldots,j_0) \tag{9}$$

for $t = \ell+1,\ldots,T$, and as

$$L = \sum_{j_{t-1},\ldots,j_0=1}^{M} \alpha_t(j_{t-1},\ldots,j_0)\, \beta_t(j_{t-1},\ldots,j_0) \tag{10}$$

for $t = 1,\ldots,\ell$. Equation (6) corresponds to the case $t = T$.

3.2 Estimation of π, A and C

The complete identification of the DCMM requires the estimation of three sets of probabilities: π, A and C. This is done using an EM algorithm known in the speech recognition litterature as the Baum-Welch algorithm. Firstly, we defind the joint probability of $\ell+1$ successive hidden states. For $t = \ell,\ldots,T-1$,

$$\begin{aligned}
\epsilon_t(j_{\ell-1},\ldots,j_0,j) &= P(X_{t-\ell+1} = j_{\ell-1},\ldots,X_t = j_0, X_{t+1} = j | Y_{-f+1},\ldots,Y_T) \\
&= \frac{\alpha_t(j_{\ell-1},\ldots,j_0)\, a_{j_{\ell-1},\ldots,j_0,j}\, c^{(j)}_{y_{t-f+1},\ldots,y_{t+1}}\, \beta_{t+1}(j_{\ell-2},\ldots,j_0,j)}{L(Y_{-f+1},\ldots,Y_T)}
\end{aligned} \tag{11}$$

Similarly, for $t = 1,\ldots\ell-1$, we define the joint probability of $t+1$ successive hidden states:

$$\begin{aligned}
\epsilon_t(j_{t-1},\ldots,j_0,j) &= P(X_1 = j_{t-1},\ldots,X_t = j_0, X_{t+1} = j | Y_{-f+1},\ldots,Y_T) \\
&= \frac{\alpha_t(j_{t-1},\ldots,j_0)\, \pi_{t+1|1,\ldots,t}\, c^{(j)}_{y_{t-f+1},\ldots,y_{t+1}}\, \beta_{t+1}(j_{t-2},\ldots,j_0,j)}{L(Y_{-f+1},\ldots,Y_T)}
\end{aligned}$$

Then, we defind the joint distribution of ℓ successive hidden states. For $t = \ell, \ldots, T$,

$$\begin{aligned} \gamma_t(j_{\ell-1}, \ldots, j_0) &= P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0 | Y_{-f+1}, \ldots, Y_T) \\ &= \frac{\alpha_t(j_{\ell-1}, \ldots, j_0)\, \beta_t(j_{\ell-1}, \ldots, j_0)}{L(Y_{-f+1}, \ldots, Y_T)} \end{aligned} \tag{12}$$

Similarly, for $t = 1, \ldots, \ell - 1$, we define the joint probability of t successive hidden states:

$$\gamma_t(j_{t-1}, \ldots, j_0) = \frac{\alpha_t(j_{t-1}, \ldots, j_0)\, \beta_t(j_{t-1}, \ldots, j_0)}{L(Y_{-f+1}, \ldots, Y_T)}$$

The following relations hold between γ and ϵ. For $t = 1, \ldots, \ell - 1$,

$$\gamma_t(j_{t-1}, \ldots, j_0) = \sum_{j=1}^{M} \epsilon_t(j_{t-1}, \ldots, j_0, j)$$

and for $t = \ell, \ldots, T - 1$,

$$\gamma_t(j_{\ell-1}, \ldots, j_0) = \sum_{j=1}^{M} \epsilon_t(j_{\ell-1}, \ldots, j_0, j)$$

Using ϵ_t and γ_t we can write the reestimation formulas of π, A and C as follows. For $t = 1$,

$$\begin{aligned} \hat{\pi}_{j_0} &= P(X_1 = j_0 | Y_{-f+1}, \ldots, Y_T) \\ &= \gamma_1(j_0) \end{aligned}$$

and for $t = 2, \ldots, \ell$,

$$\hat{\pi}_{j_0 | j_{t-1}, \ldots, j_1} = \frac{\gamma_t(j_{t-1}, \ldots, j_0)}{\gamma_{t-1}(j_{t-1}, \ldots, j_1)} \tag{13}$$

The high-order transition probabilities between hidden states are computed as

$$\hat{a}_{j_{\ell-1}, \ldots, j_0, j} = \frac{\sum_{t=\ell}^{T-1} \epsilon_t(j_{\ell-1}, \ldots, j_0, j)}{\sum_{t=\ell}^{T-1} \gamma_t(j_{\ell-1}, \ldots, j_0)} \tag{14}$$

and the high-order transitions between observations are obtained as

$$\hat{c}^{(j_0)}_{i_f, \ldots, i_0} = \frac{\sum_{t=1}^{T} \sum_{j_{\ell-1}=1}^{M} \cdots \sum_{j_1=1}^{M} \gamma_t(j_{\ell-1}, \ldots, j_0)}{\sum_{t=1}^{T} \sum_{j_{\ell-1}=1}^{M} \cdots \sum_{j_1=1}^{M} \gamma_t(j_{\ell-1}, \ldots, j_0)} \tag{15}$$

$$\begin{matrix} \\ Y_{t-f}=i_f \ldots Y_t=i_0 \end{matrix} \qquad \begin{matrix} \\ Y_{t-f}=i_f \ldots Y_{t-1}=i_1 \end{matrix}$$

In practice, the estimation of the model is achieved using iteratively the forward-backward procedures and the reestimation formulas for π, A and C.

3.3 Optimal sequence of hidden states

Given an estimation of the model, we can search the optimal sequence of hidden states which maximizes the conditional probability

$$P(X_1, \ldots, X_T | Y_{-f+1}, \ldots, Y_T)$$

or equivalently the joint probability

$$P(X_1, \ldots, X_T, Y_{-f+1}, \ldots, Y_T)$$

In speech recognition, this is known as the "global decoding problem". It can be solved through a dynamic procedure called the Viterbi algorithm (Forney, 1973). For $t = 1$ and $j_0 = 1, \ldots, M$, we define

$$\begin{aligned} \delta_1(j_0) &= P(Y_{-f+1}, \ldots, Y_1, X_1 = j_0) \\ &= \pi_{j_0} \, c^{(j_0)}_{y_{-f+1}, \ldots, y_1} \end{aligned} \tag{16}$$

For $t = 2, \ldots, \ell$ and $j_{t-1}, \ldots, j_0 = 1, \ldots, M$,

$$\begin{aligned} \delta_t(j_{t-1}, \ldots, j_0) &= \max_{j_{t-1}, \ldots, j_1} P(Y_{-f+1}, \ldots, Y_t, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1, X_t = j_0) \\ &= \left[\max_{j_{t-1}, \ldots, j_1} \delta_{t-1}(j_{t-1}, \ldots, j_1) \, \pi_{j_0 | j_1, \ldots, j_{t-1}} \right] c^{(j_0)}_{y_{t-f}, \ldots, y_t} \end{aligned} \tag{17}$$

and for $t = \ell + 1, \ldots, T$ and $j_{\ell-1}, \ldots, j_0 = 1, \ldots, M$,

$$\begin{aligned} \delta_t(j_{\ell-1}, \ldots, j_0) &= \max_{j_\ell, \ldots, j_1} P(Y_{-f+1}, \ldots, Y_t, X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1, X_t = j_0) \\ &= \left[\max_{j_\ell, \ldots, j_1} \delta_{t-1}(j_\ell, \ldots, j_1) \, a_{j_\ell, \ldots, j_1, j_0} \right] c^{(j_0)}_{y_{t-f}, \ldots, y_t} \end{aligned} \tag{18}$$

The optimal hidden state at time T is then determined as

$$\hat{x}_T = \arg \max_{j_{\ell-1}, \ldots, j_0 = 1, \ldots, M} \delta_T(j_{\ell-1}, \ldots, j_0)$$

For $t = T - 1, \ldots, \ell$, we obtain recursively

$$\hat{x}_t = \arg \max_{j_{\ell-1}, \ldots, j_0 = 1, \ldots, M} \delta_t(j_{\ell-1}, \ldots, j_0) \, a_{j_{\ell-1}, \ldots, j_0, \hat{x}_{t+1}}$$

and for $t = \ell - 1, \ldots, 1$

$$\hat{x}_t = \arg \max_{j_{t-1},\ldots,j_0=1,\ldots,M} \delta_t(j_{t-1}, \ldots, j_0)\, \pi_{\hat{x}_{t+1}|j_{t-1},\ldots,j_0}$$

Finally, the joint probability of the sequence of hidden states and the sequence of observations is

$$P(X_1, \ldots, X_T, Y_{-f+1}, \ldots, Y_T) = \max_{j_{\ell-1},\ldots,j_0=1,\ldots,M} \delta_T(j_{\ell-1}, \ldots, j_0)$$

3.4 Simultaneous data

Consider a set of N independent sequences of data. We note S_n the n^{th} sequence, T_n the number of data (not including the data used to initialize the visible Markov chain), X_t^n the t^{th} hidden state and Y_t^n the t^{th} observation. Since the sequences are independent, the major part of the computation can be made separately upon each sequence. The likelihood $L(S_n)$ of the n^{th} sequence is obtain by computing the forward-backward procedures and applying equations (9) and (10). The global likelihood of all data is then equal to

$$L(S_1, \ldots, S_N) = \prod_{n=1}^{N} L(S_n)$$

The reestimation formulas for π, A and C must take into account the information provided by the N sequences. Then,

$$\begin{aligned} \hat{\pi}_{j_0} &= P(X_1^n = j_0 | Y_{-f+1}^n, \ldots, Y_{T^n}) \quad , \quad \forall n \\ &= \frac{1}{N} \sum_{n=1}^{N} \gamma_1^n(j_0) \end{aligned}$$

$$\begin{aligned} \hat{\pi}_{j_0|j_{t-1},\ldots,j_1} &= P(X_t^n = j_0 | Y_{-f+1}^n, \ldots, Y_T^n, X_1^n = j_{t-1}, \ldots, X_{t-1}^n = j_1) \quad , \quad \forall n \\ &= \frac{1}{N} \sum_{n=1}^{N} \frac{\gamma_t^n(j_t, \ldots, j_0)}{\gamma_{t-1}^n(j_t, \ldots, j_1)} \end{aligned}$$

$$\begin{aligned} \hat{a}_{j_{\ell-1},\ldots,j_0,j} &= \sum_{t=1}^{T_n-1} P(X_{t+1}^n = j | X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t^n = j_0, Y_{-f+1}^n, \ldots, Y_{T_n}^n) \quad , \quad \forall n \\ &= \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n-1} \frac{\epsilon_t^n(j_{\ell-1}, \ldots, j_0, j)}{\gamma_t^n(j_{\ell-1}, \ldots, j_0)} \end{aligned}$$

$$\hat{c}^{(j_0)}_{i_f \ldots i_0} = P(Y^n_t = i_0 | Y^n_{-f+1}, \ldots, Y^n_{t-f} = i_f, \ldots, Y^n_{t-1} = i_1, Y^n_{t+1}, \ldots, Y^n_T, X^n_t = j_0) \quad , \quad \forall n$$

$$= \frac{\displaystyle\sum_{\substack{t=1 \\ Y^n_{t-f}=i_f \ldots Y^n_t = i_0}}^{T} \sum_{j_{\ell-1}=1}^{M} \cdots \sum_{j_1=1}^{M} \gamma^n_t(j_{\ell-1}, \ldots, j_0)}{\displaystyle\sum_{\substack{t=1 \\ Y^n_{t-f}=i_f \ldots Y^n_{t-1}=i_1}}^{T} \sum_{j_{\ell-1}=1}^{M} \cdots \sum_{j_1=1}^{M} \gamma^n_t(j_{\ell-1}, \ldots, j_0)}$$

The hidden states are obtained by running the Viterbi algorithm separately upon each sequence. If $P(S(X_n), S(Y_n))$ denotes the joint probability of hidden states and observations of the n^{th} sequence, the global joint probability is equal to

$$P(S(X_1), S(Y_1); \ldots; S(X_N), S(Y_N)) = \prod_{n=1}^{N} P(S(X_n), S(Y_n))$$

because of the independence of each sequence.

4 Modeling of π, A and C

A DCMM with M hidden states, an order ℓ hidden chain and an order f visible chain has $\sum_{g=0}^{\ell-1} M^g(M-1)$ independent parameters for π, $M^\ell(M-1)$ independent parameters for the transition matrix A, and $MK^f(K-1)$ independent parameters for the M transition matrices C. When ℓ or f is large, the total number of parameters can become too large to be estimated. In this Section, we propose to solve this problem through the modeling of π, A and C. Note that results concerning π and A can also be used to reduce the number of parameters of a Hidden Markov Model.

4.1 Modeling of π

Note first that when there is only one sequence of data, π is generally deterministic: there is a probability one to observe a particular state at each time $t = 1, \ldots, \ell$, and zero for the other states. In this case, the number of independent parameters for π is equal to zero and there is no need for a modeling.

In the general case, when there are two or more simultaneous time-series, the number of parameters for the distributions π can become unnecessarily large. We consider then two modelings of π. The first is based on an assumption of independence between the first ℓ hidden states, and the second is computed from the transition matrix A.

If we make the assumption that the first ℓ hidden states are independent one from another, we can replace each distribution π with a uniform distribution. This solution does not add any extra parameter for π what is very convenient. On the other side, it is difficult to justify the independence between the first ℓ hidden states, when there is an order ℓ dependence between the following hidden states.

A better solution is to suppose that the first hidden states follow a transition process similar to the one of the next states, that is conform to the transition matrix A. This matrix can then be used to compute an approximation of π. For the first hidden state distribution,

$$\begin{aligned}
\hat{\pi}_{j_0} &= P(X_1 = j_0 | Y_{-f+1}, \ldots, Y_T) \\
&\approx P(X_t = j_0 | Y_{-f+1}, \ldots, Y_T) \quad , \forall t = \ell + 1, \ldots, T \\
&\approx \frac{1}{M^\ell} \sum_{j_\ell, \ldots, j_1 = 1}^{M} P(X_t = j_0 | Y_{-f+1}, \ldots, Y_T, X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1) \\
&= \frac{1}{M^\ell} \sum_{j_\ell, \ldots, j_1 = 1}^{M} a_{j_\ell, \ldots, j_0}
\end{aligned}$$

and for $k = 2, \ldots, \ell$,

$$\begin{aligned}
&\hat{\pi}_{j_0 | j_{k-1}, \ldots, j_1} \\
&= P(X_k = j_0 | Y_{-f+1}, \ldots, Y_T, X_1 = j_{k-1}, \ldots, X_{k-1} = j_1) \\
&\approx P(X_t = j_0 | Y_{-f+1}, \ldots, Y_T, X_{t-k+1} = j_{k-1}, \ldots, X_{t-1} = j_1) \quad , \forall t = \ell + 1, \ldots, T \\
&\approx \frac{1}{M^{\ell-k+1}} \sum_{j_\ell, \ldots, j_k = 1}^{M} P(X_k = j_0 | Y_{-f+1}, \ldots, Y_T, X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1) \\
&= \frac{1}{M^{\ell-k+1}} \sum_{j_\ell, \ldots, j_k = 1}^{M} a_{j_\ell, \ldots, j_0}
\end{aligned}$$

This second solution is better than the first, since it keeps a dependence relation between all hidden states. Moreover, since π depends on A only, it does not add any supplementary parameter to the model.

4.2 Modeling of A

When the order of the hidden relation is greater than 1, it is possible to replace the transition matrix A by a modeling. Among the class of Markov chain modelings, the Mixture Transition Distribution model (MTD) introduced by Raftery (1985) is particularly interesting. The idea is to consider separately the effect of each lag upon the period t:

$$\begin{aligned}\hat{a}_{j_\ell,\ldots,j_0} &= P(X_t = j_0 | X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1) \\ &\approx \sum_{g=1}^{\ell} \lambda_g P(X_t = j_0 | X_{t-g} = j_g) \\ &= \sum_{g=1}^{\ell} \lambda_g q_{j_g j_0}\end{aligned}$$

where $j_0, \ldots, j_\ell \in \{1, \ldots, M\}$, $Q = [q_{j_g j_0}]$ is a $M \times M$ row transition matrix, and $\lambda = \{\lambda_1, \ldots, \lambda_\ell\}$ is a vector of lag parameters, with $\sum_{g=1}^{\ell} \lambda_g = 1$.

The MTD model is far more parsimonious than the corresponding whole parameterized Markov chain since, in addition to the $M(M-1)$ independent parameters of the transition matrix Q, it requires only one additional parameter for each supplementary lag. It is also possible to use a different transition matrix Q to represent the relation between each lag and the period t. The resulting model (MTDg) can handle a greather set of situations, but it is less parsimonious. The MTD model is estimated by maximizing the log-likelihood

$$\log(L) = \sum_{j_\ell,\ldots,j_0=1}^{M} n_{j_\ell,\ldots,j_0} \log \left(\sum_{g=1}^{\ell} \lambda_g q_{j_g j_0} \right)$$

where $n_{j_\ell,\ldots,j_0}$ is the number of sequences of the form

$$X_{t-\ell} = j_\ell, \ldots, X_t = j_0 \tag{19}$$

in the data.

The estimation of the MTD model requires the knowledge of the quantities $n_{j_\ell,\ldots,j_0}$. In the case of the DCMM, these quantities are unknown, but they can be estimated. At time t, the probability to observe (19) is given by $\epsilon_{t-1}(j_\ell, \ldots, j_0)$. We can then approximate the number of appearances of this sequence throughout the whole set of data as

$$\hat{n}_{j_\ell,\ldots,j_0} = \sum_{t=\ell}^{T-1} \epsilon_t(j_\ell, \ldots, j_0)$$

Note that the quantities $\hat{n}_{j_\ell,\ldots,j_0}$ can be non-integer. With $\hat{n}_{j_\ell,\ldots,j_0}$ known, the MTD model can be estimated using a constrained optimization procedure. Raftery & Tavaré (1994) used a software called MTD, but the method introduced in Berchtold (1999b) is more convenient since it does not require the use of any external optimization procedure and can be implemented on any computing platform. A similar approach was used once by Schimert (1992) in the context of a high-order Hidden Markov Model.

4.3 Modeling of C

The modeling of C is achieved through the same principle used for A. When the order of the visible chain is greater than 1, each transition matrix C can be replaced by a MTD model. For each hidden state $j_0 = 1, \ldots, M$, the number of transitions of the form

$$Y_{t-f} = i_f, \ldots, Y_t = i_0$$

given $X_t = j_0$ is estimated as

$$\hat{n}^{(j_0)}_{i_f,\ldots,i_0} = \sum_{t=1}^{T} P(X_t = j_0)\, \psi_t(i_f, \ldots, i_0)$$

where

$$P(X_t = j_0) = \sum_{j_\ell,\ldots,j_1=1}^{M} \epsilon_t(j_\ell, \ldots, j_0)$$

and

$$\psi_t(i_f, \ldots, i_0) = \begin{cases} 1 & \text{, if } Y_{t-f} = i_f, \ldots, Y_t = i_0 \\ 0 & \text{, otherwise} \end{cases}$$

Using the MTD model, each transition matrix C requires the estimation of only $K(K-1)+f-1$ independent parameters, what is far more parsimonious than the $K^f(K-1)$ parameters of the whole parameterized transition matrix.

5 Applications

In this Section, we analyze several sets of data showing different fields of application of the Double Chain Markov Model. Comparisons between models are carrying out using the Bayesian Information Criterion (BIC), which is defined as

$$BIC = -2\log(L) + p\log(T)$$

where L is the likelihood of the model, p is the number of independent parameters and T is the number of components in the likelihood. The model with the lowest BIC is chosen. The application of this criterion to Markovian models was discussed by Katz (1981). According to the convention established by Bishop et al. (1975) we do not take into account the parameters estimated to zero.

5.1 DNA analysis

DNA sequences can be decomposed into an alphabet of 4 bases {A, C, G, T}. Since these 4 bases do not occur randomly inside a gene, it is of particular interest to find a model giving the structure of a sequence. Here, we study a particular binary decomposition of the 4 letter alphabet, the purine-pyrimidine alphabet. Each base is recoded as either purine ({A, G}) or pyrimidine ({C, T}). We consider the mouse αA-crystallin gene previously analyzed by Avery (1987) and Raftery & Tavaré (1994). This is a length 1307 sequence, but we dropped the first 5 data in order to have exactly 1302 components in the log-likelihood of each model. The complete data appear in Table 7 of Raftery & Tavaré (1994).

Raftery & Tavaré (1994) analyzed these data using Markov chains and found that a second-order Mixture Transition Distribution model gives the best results. We tried to improve their results by the use of either a Hidden Markov Model or a Double Chain Markov Model. Table 1 and 2 report our results.

According to BIC, and in spite of a greather number of parameters (6 instead of 3), the DCMM 2 (1;1) fits the data better than the MTD 2. This result is obtained through a significative improvement of the log-likelihood. The different models appearing in Tables 1 and 2 show the interest and the consequences of the use of a MTD model to replace a high-order Markov chain inside a HMM or a DCMM. According to the theory, a model using a MTD modeling achieves a lower log-likelihood than the corresponding full parameterized model, but since it has a smaller number of parameters, it can often obtain a better BIC value.

5.2 Song of the wood pewee

We consider a time-series of length 1327, with three possible values corresponding to the three distinct phrases of the wood pewee song. The complete time-series appears in Table 4. These data were originally described by Craig (1943) and reanalyzed in Chatfield & Lemon (1970), Bishop et al. (1975), Raftery & Tavaré (1994) and Berchtold (1999a). Among the data, two patterns are highly represented. The pattern *1312* occurs 260 times and the pattern *112* occurs 40 times. Our results are summarized in Table 3. Note that we dropped the first data of the sequence in order to have exactly 1323 components in the log-likelihood of each model. Moreover, we did not consider the parameters equal to zero for the computation of BIC. This explains why some of our results are different from those of Raftery & Tavaré.

The data exhibit a clear relation between successive observations, so neither the independence model nor the Hidden Markov Model achieve good results. Among the full parameterized Markov chains, the best model is the fourth order chain, and even if it doesn't appear in Table 3, the use

Table 1. Different modelings of the purine-pyridine decomposition of a DNA sequence, part 1. MC f is a Markov chain of order f. MTD f is a Mixture Transition Distribution model of order f. HMM x (ℓ) is a x states Hidden Markov Model of order ℓ. A "M" denotes a MTD modeling.

Model	Number of parameters	Log-likelihood	BIC
Independence	1	-901.86	1810.9
MC 1	2	-889.25	1792.8
MC 2	4	-884.73	1798.1
MC 3	8	-878.20	1813.8
MC 4	16	-869.36	1853.5
MC 5	32	-858.50	1946.5
MTD 2	**3**	**-884.89**	**1791.3**
MTD 3	4	-884.17	1797.0
MTD 4	5	-880.84	1797.5
MTD 5	6	-880.73	1804.5
HMM 2 (1)	4	-888.68	1806.0
HMM 2 (M2)	5	-883.51	1802.9
HMM 2 (2)	6	-883.39	1809.8
HMM 2 (M3)	6	-881.82	1806.7
HMM 2 (3)	10	-867.50	1806.7
HMM 2 (4)	17	-862.09	1846.1
HMM 3 (1)	9	-881.68	1827.9
HMM 3 (M2)	10	-881.38	1834.5
HMM 3 (2)	17	-870.76	1863.4
HMM 4 (1)	12	-875.35	1836.8

of a fifth or sixth order model doesn't improve the results. Raftery & Tavaré tried also different non-Markovian modelings and, using the pattern structure of the time-series, their best model achieved a BIC of 827 ("Pattern" in Table 3). In Berchtold (1999b) we used the Mixture Transition Distribution model with a different matrix for each lag (MTDg in Table 3), but it did not beat the fully parameterized Markov chains. So, we tried different DCMM, and among them a model with two hidden states and second order transition matrices between both hidden states and observations proved to be better than any other modeling ("DCMM 2 (2;2)" in Table 3). The use of higher order dependences did not improve the results. It is interesting to note that our final model achieves a better log-likelihood than the fourth order fully parameterized Markov chain does, with less parameters.

We give hereafter the estimated parameters of the DCMM 2 (2;2) model. The distribution of the first hidden state is $\pi = (0 \quad 1)$, what means that the process starts in state 2 with probability

Table 2. Different modelings of the purine-pyridine decomposition of a DNA sequence, part 2. DCMM x $(\ell;f)$ is a x states Double Chain Markov Model with a hidden dependence of order ℓ and a visible dependence of order f. A "M" denotes a MTD modeling.

Model	Number of parameters	Log-likelihood	BIC
DCMM 2 (1;1)	**6**	**-873.71**	**1790.5**
DCMM 2 (2;1)	8	-872.71	1802.8
DCMM 2 (1;M2)	8	-873.57	1804.5
DCMM 2 (1;2)	10	-863.32	1798.4
DCMM 2 (M2;M2)	9	-867.98	1800.5
DCMM 2 (2;M2)	10	-860.01	1791.7
DCMM 2 (2;2)	12	-859.31	1804.7
DCMM 2 (1;M3)	10	-863.53	1798.8
DCMM 2 (2;M3)	12	-859.18	1804.4
DCMM 2 (M3;M3)	12	-866.17	1818.4
DCMM 3 (1;1)	10	-866.29	1804.3
DCMM 3 (2;1)	16	-850.60	1815.9
DCMM 3 (1;2)	17	-856.82	1835.6
DCMM 3 (2;2)	21	-848.71	1848.0

one. The conditional distribution of the second hidden state given the first is

$$\pi_{2|1} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

This indicates that the second hidden state is the state 2 with probability one, whatever the first hidden state is. As noted in Section 4.1, since we have here only one sequence of data, the first two hidden states are deterministic. The transition matrix between hidden states is (in reduced form)

$$A = \begin{array}{cc|cc} & & \multicolumn{2}{c}{X_t} \\ X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 0.9618 & 0.0382 \\ 2 & 1 & 0.5171 & 0.4829 \\ 1 & 2 & 0.9676 & 0.0324 \\ 2 & 2 & 0.0097 & 0.9903 \end{array}$$

When the first state appears two times consecutively, it has a great probability (0.9618) to appear one more time. We observe the same behavior for the second hidden state. The third row indicates that the second state can occasionaly appear inside a subsequence of several consecutive first state, what can be verified in Table 4. The transition matrix between observations given the first hidden state is (in reduced form)

Table 3. Different modelings of the wood pewee song. HMM 2 (1) is a two states first-order Hidden Markov Model, MC f is an order f Markov Chain, "Pattern" is the best modeling found by Raftery & Tavaré (1994), MTDg f is an order f MTD model with a different matrix for each lag, DCMM x $(\ell;f)$ is a x states Double Chain Markov Model, where ℓ is the order of the hidden chain, f the order of the visible chain, and where "M" denotes a MTD modeling and "Mg" denotes a MTDg modeling.

Model	Number of parameters	Log-likelihood	BIC
Independence	2	-1349.4	2713.3
HMM 2 (1)	7	-1086.4	2223.2
MC 1	5	-694.1	1424.2
MC 2	9	-368.6	801.9
MC 3	14	-354.0	808.6
MC 4	19	-315.8	768.3
Pattern	4	-399.1	827.0
MTDg 2	9	-486.4	1037.4
MTDg 3	9	-484.0	1032.7
DCMM 2 (1;2)	17	-367.5	857.2
DCMM 2 (1;M2)	7	-384.1	818.4
DCMM 2 (1;Mg2)	11	-382.1	843.2
DCMM 2 (2;2)	**17**	**-305.4**	**733.0**
DCMM 2 (2;M2)	9	-383.8	832.2
DCMM 3 (1;2)	15	-344.0	795.8
DCMM 3 (2;2)	23	-304.6	774.5

$$C_1 = \begin{array}{cc|ccc} & & & Y_t & \\ Y_{t-2} & Y_{t-1} & 1 & 2 & 3 \\ \hline 1 & 1 & 0 & 0.6424 & 0.3576 \\ 2 & 1 & 0 & 0.0639 & 0.9361 \\ 3 & 1 & 0.0175 & 0.9703 & 0.0122 \\ 1 & 2 & 0.9928 & 0 & 0.0072 \\ 2 & 2 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 \end{array}$$

Finally, the transition matrix between observations given the second hidden state is (in reduced form)

$$C_2 = \begin{array}{cc|ccc} & & & Y_t & \\ Y_{t-2} & Y_{t-1} & 1 & 2 & 3 \\ \hline 1 & 1 & 0.0970 & 0.8163 & 0.0867 \\ 2 & 1 & 0.6306 & 0.1445 & 0.2249 \\ 3 & 1 & 0.1169 & 0.8831 & 0 \\ 1 & 2 & 0.9874 & 0 & 0.0126 \\ 2 & 2 & 0.25 & 0.75 & 0 \\ 3 & 2 & 1 & 0 & 0 \\ 1 & 3 & 0.9752 & 0.0248 & 0 \\ 2 & 3 & 1 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 \end{array}$$

Note that since phrase 3 is never followed by itself, the last row of C_1 and C_2 was not computed. The pattern *1312* appears very clearly in matrix C_1. If we consider only subsequences 131, 312, 121 and 213, we have

$$C_1' = \begin{array}{cc|ccc} & & & Y_t & \\ Y_{t-2} & Y_{t-1} & 1 & 2 & 3 \\ \hline 1 & 1 & - & - & - \\ 2 & 1 & - & - & 0.9361 \\ 3 & 1 & - & 0.9703 & - \\ 1 & 2 & 0.9928 & - & - \\ 2 & 2 & - & - & - \\ 3 & 2 & - & - & - \\ 1 & 3 & 1 & - & - \\ 2 & 3 & - & - & - \\ 3 & 3 & - & - & - \end{array}$$

The product of these four probabilities gives 0.9018. So, once one of these transitions occurs, there is a very high probability to create a sequence of the form *...13121312...*. The other elements of this matrix correspond to short sequences of observations appearing only marginally in the data.

The transition matrix C_2 corresponding to the second hidden state is somewhat more difficult to analyze. The pattern *112* has a good probability of appearance, but with a probability of 0.5083 it is less dominant than *1312* in C_1:

$$C_2' = \begin{array}{cc|ccc} & & & Y_t & \\ Y_{t-2} & Y_{t-1} & 1 & 2 & 3 \\ \hline 1 & 1 & - & 0.8163 & - \\ 2 & 1 & 0.6306 & - & - \\ 3 & 1 & - & - & - \\ 1 & 2 & 0.9874 & - & - \\ 2 & 2 & - & - & - \\ 3 & 2 & - & - & - \\ 1 & 3 & - & - & - \\ 2 & 3 & - & - & - \\ 3 & 3 & - & - & - \end{array}$$

The pattern *1312* also appears in this matrix:

$$C_2'' = \begin{array}{cc|ccc} & & & Y_t & \\ Y_{t-2} & Y_{t-1} & 1 & 2 & 3 \\ \hline 1 & 1 & - & - & - \\ 2 & 1 & - & - & 0.2249 \\ 3 & 1 & - & 0.8831 & - \\ 1 & 2 & 0.9874 & - & - \\ 2 & 2 & - & - & - \\ 3 & 2 & - & - & - \\ 1 & 3 & 0.9752 & - & - \\ 2 & 3 & - & - & - \\ 3 & 3 & - & - & - \end{array}$$

but with a probability as low as 0.1912, it is difficult to observe even only one complete realization of this pattern.

The model roughly decomposes the whole sequence into three parts. This is shown in Table 4, where the optimal hidden state sequence is given on the first row and the corresponding observations are on the second row. The model is in state 2 during the first 40 observations, then it switches to state 1 for the following 1028 observations, and finally goes back to state 2 for the last 259 observations. We must note that this decomposition into three parts is not perfect. State 2 appears sometimes in the median part of the time-series, for instance at time 180 and 229, what indicates a rupture in the *1312* pattern. Conversely, the subsequence *1312* can appear briefly when the process is in state 2, particularly in the beginning of the third part of the time-series. Nevertheless, such events are rare.

This example shows clearly the interest of the Double Chain Markov Model. It handles non-homogeneous data very well and specific patterns like *1312* and *112* can be traced back in the transition matrices. Moreover, the hidden states provide a tool to decompose the whole set of observations into a finite number of categories.

Table 4. Hidden states for the wood pewee song. The first row gives the optimal sequence of hidden states for the DCMM 2 (2;2). The second row gives the corresponding observations of the wood pewee song. Note that since we dropped the first four observations, there are no corresponding hidden states.

2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 1 1	2 1 1 2 1 1 2 1 1 2	1 1 2 1 1 2 1 1 3 1	2 1 2 1 1 2 1 1 2 1	3 1 2 1 3 1 2 1 2 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	2 1 2 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
3 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	2 1 1 2 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 1 3 1 2 1 3 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 2	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
2 1 3 1 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 1	2 1 3 1 2 1 3 1 2 1	2 1 3 1 2 1 3 1 2 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 2 1	2 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
3 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 2 1	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 2 1 3 1 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 3 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
2 1 3 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	2 1 3 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 3 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 2 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
2 1 3 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 1 3	1 1 2 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
3 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	3 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 2 1 3 1	2 1 3 1 2 1 3 1 3 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
2 1 3 1 2 3 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 1 2 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	2 1 3 1 2 1 3 1 2 1
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
3 1 2 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 2 1 3	1 2 1 3 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 2 1 3 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 2	1 3 1 2 1 3 1 2 1 3
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 3 1 2 1 3	1 2 1 3 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 2	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 2 1 1 1
1 3 1 2 1 3 1 2 1 3	1 2 1 2 1 3 1 2 1 3	1 2 1 3 1 3 1 2 1 1	2 1 3 1 2 1 2 1 3 1	2 1 3 1 2 1 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 2 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
1 3 1 2 1 3 1 2 1 3	1 2 1 1 3 1 2 1 3 1	2 1 3 1 2 1 3 1 2 1	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 2 1 3 1 2
1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2
1 3 1 2 1 3 1 2 1 3	1 2 1 2 1 3 1 2 1 1	2 1 3 1 1 2 1 1 2 1	1 3 2 1 1 2 1 1 3 1	2 1 2 1 3 1 2 1 3 1
2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2
2 1 3 1 2 1 3 1 2 1	2 1 3 1 2 1 3 1 2 1	1 2 1 3 1 2 1 3 1 2	1 2 1 3 1 2 1 3 1 2	1 3 1 2 1 1 2 1 3 1
2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2
1 2 1 3 1 2 1 3 1 2	1 1 3 1 2 1 1 2 1 1	2 1 1 2 1 1 2 1 1 2	1 1 1 2 1 3 1 2 1 1	2 1 1 2 1 3 1 2 1 3
2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2
1 2 1 1 2 1 1 2 1 3	1 2 1 3 1 2 1 2 1 1	3 1 2 1 2 1 2 1 1 2	1 3 1 2 1 2 1 2 1 1	2 3 1 1 2 1 3 1 2 1
2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2
1 2 1 3 1 2 1 1 2 1	3 1 2 1 2 1 1 1 2 1	1 2 1 3 1 2 1 2 1 3	1 2 1 2 1 3 1 2 1 1	2 1 1 1 1 2 1 1 2 1
2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2		
1 2 1 1 2 1 1 2 1 1	2 1 1 2 1 1 2 1 1 2	1 1 1 2 1 1 2		

5.3 Behavior of young monkeys

A Double Chain Markov Model can be used to explain the behavior of young rhesus monkeys. The data, provided by Gene Sackett, were collected using the following procedure: Young monkeys separated from their mother were socialized with other young monkeys during 5 minute periods and the behavior of one monkey (focal subject) and its relations with the other monkeys (interactors) were observed. The role of the focal subject in the interaction with the other subjects (Nonsocial,

Initiate with contact, No response without contact, ...) was recorded along with the behavior of both the focal and the interactor monkeys. Only four behaviors were considered (Passive, Explore, Fear/Disturb, Play), the others like "Sex" and "Aggression" being not present or appearing only marginally in the data. More details about the procedure used to collect the data and traditional analysis can be found for instance in Worlein & Sackett (1997) and in Novak & Sackett (1997). It must be noted that since these data were obtained by observing and interpreting the behavior of young monkeys, they are subjective and can possibly include mistakes.

Our goal was to find a model explaining the behavior of the focal subject in function of external influences (the interactors), but without explicitly putting them into the model. Therefore, a hidden model seemed appropriate, but since the sequence of successive behaviors of monkeys are not completely random, we rejected the HMM because of the conditional independence hypothesis implied by this model. The DCMM seemed then a better choice. Since we are interested only in the transition process between the different behaviors of the focal subject, we did not consider neither the duration of each behavior nor the role of the focal subject in the interaction.

We considered data from two different monkeys. The first was observed during 15 sessions ranging from 32 to 109 days of age. The second monkey was observed during 11 sessions ranging from 62 to 127 days of age. Table 5 presents our results for the analysis of the first monkey.

The independence model is clearly rejected and, among the Markov chains, the MTD 3 model obtains the best result. In spite of a larger number of parameters, the best overall result is achieved by the DCMM 2 (1;1) with two hidden states and first-order transition matrices for both hidden and visible chains. The distribution of the first hidden state is $\pi = (0.3245 \quad 0.6755)$, the transition matrix between hidden states is

$$A = \begin{pmatrix} 0.9737 & 0.0263 \\ 0.0679 & 0.9321 \end{pmatrix}$$

and the corresponding transition matrices between the four possible behaviors (Passive, Explore, Fear/Disturb, Play) are

$$C_1 = \begin{pmatrix} 0.1020 & 0.4761 & 0.0840 & 0.3379 \\ 0.6239 & 0.1298 & 0 & 0.2463 \\ 0.4568 & 0.1948 & 0 & 0.3484 \\ 0.2314 & 0.1195 & 0.0224 & 0.6267 \end{pmatrix}$$

$$C_2 = \begin{pmatrix} 0.0709 & 0.1900 & 0.7391 & 0 \\ 0.2057 & 0.2963 & 0.4980 & 0 \\ 0.6801 & 0.1561 & 0.1638 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Table 5. Different models for the first monkey. MC f is a Markov chain of order f. MTD f is a Mixture Transition Distribution model of order f. MTDg f is a Mixture Transition Distribution model of order f with a different transition matrix for each lag. DCMM x (ℓ;f) is a x states Double Chain Markov Model with a hidden dependence of order ℓ and a visible dependence of order f. "Mg" denotes a MTDg modeling. A "*" indicates that the π parameters were derived from the transition matrix A. The log-likelihood of each model has exactly 467 components.

Model	Number of parameters	Log-likelihood	BIC
Independence	3	-557.03	1132.5
MC 1	12	-433.51	940.8
MC 2	41	-380.34	1012.7
MC 3	64	-316.22	1025.8
MTD 2	13	-426.55	933.0
MTDg 2	23	-401.86	945.1
MTD 3	13	-417.48	914.9
MTDg 3	22	-390.89	917.0
DCMM 2 (1;1)	**19**	**-394.52**	**905.8**
DCMM 2 (1;2)	46	-337.32	957.4
DCMM 2 (1;Mg2)	42	-393.96	1046.1
DCMM 3 (1;1)	25	-379.79	913.2
DCMM 3 (2;1)	34	-378.80	966.6
DCMM 3* (2;1)	27	-384.41	934.8

The transition matrix corresponding to the second hidden state (C_2) is particularly interesting. It can be decomposed into two separate sets of behaviors, {Passive, Explore, Fear/Disturb} and {Play}, and it is impossible to go from one set to another. So, this two-state model identify really three different situations. In the first, corresponding to the transition matrix C_1, the young monkey can switch from any behavior to any other behavior in at most two steps. We rename this situation state "a". In the second situation ("b"), the monkey can have only one of the first three behaviors, excluding playing, and in the third situation ("c") "Play" is the only behavior. The state decomposition of the data is given in Table 6.

The 15 sessions can be decomposed in two parts. During the first 7 sessions, only states "a" and "b" occur. A look at the original data shows that the behavior "Play" is rare during this period. On the other hand, during the last 8 sessions, state "b" is replaced by state "c". This shows an important change in the behavior of this subject, going from an attitude dominated by "Passive" and "Fear/Disturb" to a behavior mostly turned to "Play".

The question was then to know if, observed in the same conditions, the second monkey could

Table 6. Hidden states of the DCMM 2 (1;1) for the first monkey. State "a" can represent all four behaviors, state "b" can represent only "Passive", "Explore and "Fear/Disturb", and state "c" represents exclusively "Play". The age is given in days.

Session	Age	States
1	32	bbbbb
2	34	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
3	52	bbbbbbbbbbbaaaabbbbbb
4	54	bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
5	59	bbbaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
6	66	aaaaaaaaaaaaaaaaaaaaaaaaaaaa
7	68	bbbb
8	76	aaaaaaaaaaaaaaaaaaaaaaaa
9	80	cccccaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
10	83	ccccccccccccccccccccccccaaccccccccccccccccc
11	88	ccccccccaaaaaaaaaaaaaaaaaaaaaaaaaa
12	90	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
13	102	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
14	104	ccccccccccccccaaaccccccccccccccccccccccc
15	109	ccccccccccccaaaaaaaaaaaaaaaaaaaaaa

be represented by a similar model. The DCMM 2 (1;1) for the second monkey has parameters $\pi = (0.9091 \quad 0.0909)$,

$$A = \begin{pmatrix} 0.9873 & 0.0127 \\ 0 & 1 \end{pmatrix}$$

and

$$C_1 = \begin{pmatrix} 0.1514 & 0.1505 & 0 & 0.6981 \\ 0.3934 & 0.1516 & 0 & 0.4549 \\ 1 & 0 & 0 & 0 \\ 0.1301 & 0.1361 & 0 & 0.7337 \end{pmatrix}$$

$$C_2 = \begin{pmatrix} 0.1832 & 0.0949 & 0.7219 & 0 \\ 0.6696 & 0 & 0.3304 & 0 \\ 0.8095 & 0 & 0.1905 & 0 \\ 0 & 0.0236 & 0 & 0.9764 \end{pmatrix}$$

Contrarily to what was observed for the first monkey, the second transition matrix is not perfectly separated into two mutually exclusive situations. Nevertheless, the global structure of this model is close to the one we found for the first monkey. Moreover, a look at the data shows that the only behavior occuring when C_2 is active is "Play", except for one appearance of "Passive" and one

appearance of "Explore" at the end of session 9 for a total duration of 6 seconds. Considering the subjectivity of these data, we tried to recode these events as "Play". The last row of C_2 was then rewritten as $(0 \quad 0 \quad 0 \quad 1)$, and adopting the states "a", "b" and "c" previously used for the first monkey, we obtained the decomposition given in Table 7.

Table 7. Hidden states of the DCMM 2 (1;1) for the second monkey. State "a" can represent all four behaviors, state "b" can represent only "Passive", "Explore and "Fear/Disturb", and state "c" represents exclusively "Play". The age is given in days.

Session	Age	States
1	62	bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
2	85	aaaaaaaaaaaaaaaaaaaaaaaaa
3	96	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
4	98	aaaaaaaaaaaaaaa
5	105	aaaaaaaaaaaaaaaaaaaa
6	111	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
7	113	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
8	118	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
9	120	aaaaaccccccccccccccccccccc
10	125	aaaaaaaaaaaaaaaccccccccccccc
11	127	aaaaaaaaaaaaaaaaaaa

Compared to the first monkey, this second subject is a lot more often in a situation where he can switch from any behavior to any other behavior ("a") and he is only one time in state "b" (session 1). This can be explain by considering the respective age of each subject. The second subject is older and its data correspond roughly to the second part of the data of the first monkey. So it is perfectly coherent to not observe state "b" after the first session of the second monkey, since this state was observed only during the first sessions for the first monkey. To confirm that, we recomputed the DCMM 2 (1;1) model for the first monkey, using only sessions 8 to 15. We obtained $\pi = (0.3789 \quad 0.6211)$,

$$A = \begin{pmatrix} 0.9843 & 0.0157 \\ 0.0219 & 0.9781 \end{pmatrix}$$

and

$$C_1 = \begin{pmatrix} 0.1146 & 0.3648 & 0 & 0.5206 \\ 0.4612 & 0 & 0 & 0.5388 \\ 1 & 0 & 0 & 0 \\ 0.2187 & 0.1130 & 0 & 0.6683 \end{pmatrix}$$

$$C_2 = \begin{pmatrix} 0.0846 & 0.4538 & 0.0598 & 0.4018 \\ 0.9055 & 0.0945 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0.0354 & 0.0152 & 0 & 0.9494 \end{pmatrix}$$

Except for the first row of C_2, this latter model is closer to the model obtained for the second monkey than the model using all 15 sessions of the first monkey. Most of the remaining differences can be attributed to the small number of data and to the fact that "Fear/Disturb" appears only once here. Nevertheless, the same general structure is found once again, with a high probability (0.9494) to stay in the behavior "Play" when matrix C_2 is active. Given this, the DCMM 2 (1;1) fits well both sets of data and it is clear that the global behavior of the two monkeys is similar, even if there are some differences in the parameters due to the particularities of each focal subject and its interactors.

6 Conclusion

In this article, we developped a fully Markovian model for the representation of non-homogeneous time-series called the Double Chain Markov Model (DCMM). This model can be viewed as an extension of both Markov chains in random environment and Hidden Markov Models. We considered the inclusion of high-order dependences between hidden and/or visible events and we provided the complete derivation of the estimation algorithms. Since a fully parameterized DCMM can have a very large number of parameters, we considered also the modeling of high-order dependences and we showed that the Mixture Transition Distribution model (MTD) is very useful in this context. A large part of this article is devoted to applications showing the use of the DCMM for the analysis of different types of data. In particular, it proved to be able to fit very well data presenting repetitive patterns, and the analysis of the corresponding hidden states provides a method to decompose and classify such data.

Acknowledgements

This research was supported by a grant from the Swiss National Science Foundation and by Office of Naval Research grant no. N00014-96-1-0192. I would like to thank Adrian Raftery for his very helpful comments and Gene Sackett for providing the young monkeys data.

References

Avery, P. J. (1987) The analysis of intron data and their use in the detection of short signals. *Journal of Molecular Evolution*, 26, 335-340.

Berchtold, A. (1999a) The Double Chain Markov Model. Technical Report 348, Department of Statistics, University of Washington. To appear in *Communications in Statistics: Theory and Methods*, 28 (11).

Berchtold, A. (1999b) Estimation of the Mixture Transition Distribution Model. Technical Report 352, Department of Statistics, University of Washington.

Bishop, Y. M. M., S. E. Fienberg, P. W. Holland (1975) *Discrete Multivariate Analysis*. MIT Press, Cambridge.

Chatfield, C., R. E. Lemon (1970) Analysing sequences of behavioral events. *Journal of Theoretical Biology*, 29, 427-445.

Cogburn, R. (1984) The Ergodic Theory of Markov Chains in Random Environments. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 66, 109-128.

Collins, E. J., J. M. McNamara (1998) Finite-horizon dynamic optimisation when the terminal reward is a concave functional of the distribution of the final state. *Advances in Applied Probability*, 30, 122-136.

Craig, W. (1943) *The Song of the Wood Peewee*. Albany, University of the State of New York.

Dynkin, E. B. (1965) *Markov Processes, Vol. I & II*. Springer-Verlag, Berlin.

Elliott, R. J., L. Aggoun, J. B. Moore (1995) *Hidden Markov Models: Estimation and Control*. Springer-Verlag, New York.

Forney, G. D. (1973) The Viterbi Algorithm. *Proceedings of the IEEE*, 61, 268-278.

Haslett, J., A. E. Raftery(1989) Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource (with Discussion). *Applied Statistics*, 38, 1-50.

Katz, R. W. (1981) On some criteria for estimating the order of a Markov chain. *Technometrics*, 23, 243-249.

Kemeny, J. G., J. L. Snell (1976) *Finite Markov Chains*. Springer-Verlag, New York.

KENNY, P., M. LENNIG, P. MERMELSTEIN (1990) A linear predictive HMM for vector-valued observations with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 38 (2), 220-225.

KIJIMA, M. (1997) *Markov Processes for Stochastic Modeling*. Chapman & Hall, London.

MACDONALD, I. L., W. ZUCCHINI (1997) *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall, London.

NOVAK, M. F. S. X., G. P. SACKETT (1997) Pair-Rearing Infant Monkeys (Macaca nemestrina) Using a "Rotating-Peer" Strategy. *American Journal of Primatology*, 41, 141-149.

PALIWAL, K. K. (1993) Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer. *Proceedings ICASSP*, Vol. 2, 215-218.

PORITZ, A. B. (1982) Linear predictive hidden Markov models and the speech signal. *Proceedings ICASSP*, 1291-1294.

PORITZ, A. B. (1988) Hidden Markov models: A guided tour. *Proceedings ICASSP*, Vol. 1, 7-13.

RABINER, L. R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol. 77, No 2, 257-286.

RAFTERY, A. E. (1985) A model for high-order Markov chains. *Journal of the Royal Statistical Society B*, Vol. 47, No 3, 528-539.

RAFTERY, A. E., S. TAVARÉ (1994) Estimation and Modelling Repeated Patterns in High Order Markov Chains with the Mixture Transition Distribution Model. *Applied Statistics*, Vol. 43, No 1, 179-199.

SCHIMERT, J. (1992) A high order hidden Markov model. Ph. D. thesis 40908, University of Washington, USA.

WELLEKENS, C. J. (1987) Explicit time correlation in Hidden Markov Models for speech recognition. *Proceedings ICASSP*, 384-386.

WORLEIN, J. M., G. P. SACKETT (1997) Social development in nursery-reared pigtailed macaques (Macaca nemestrina). *American Journal of Primatology*, 41, 23-35.

A Derivation of the results of Section 3

We provide in this appendix the complete derivation of the algorithms presented in Section 3. More details about their practical implementation can be found in Berchtold (1999a). Note that even if the principle of these algorithms is not new, it is interesting to give them here with some details since, in the contrary of previous publications, we consider a very general model including Markovian relations between both hidden and visible variables, and high-order dependencies. This leads to more complex derivations, especially for the first terms of the time-series.

A.1 Likelihood of the observed output sequence

The forward terms α_t are obtained as follows. For $t = 1$, equation (2) is

$$\begin{aligned}
\alpha_1(j) &= P(Y_{-f+1}, \ldots, Y_1, X_1 = j) \\
&= P(Y_1 | Y_{-f+1}, \ldots, Y_0, X_1 = j)\, P(Y_{-f+1}, \ldots, Y_0, X_1 = j) \\
&= c^{(j)}_{y_{-f+1}, \ldots, y_1}\, P(Y_{-f+1}, \ldots, Y_0, X_1 = j)
\end{aligned}$$

Since $Y_{-f+1}, \ldots, Y_0$ are independent of X_1 and the value of $Y_{-f+1}, \ldots, Y_0$ is known, we can write

$$\begin{aligned}
\alpha_1(j) &= c^{(j)}_{y_{-f+1}, \ldots, y_1}\, P(Y_{-f+1}, \ldots, Y_0)\, P(X_1 = j) \\
&= c^{(j)}_{y_{-f+1}, \ldots, y_1}\, P(X_1 = j) \\
&= c^{(j)}_{y_{-f+1}, \ldots, y_1}\, \pi_j
\end{aligned}$$

For $t = 2$, equation (3) is

$$\begin{aligned}
\alpha_2(j_1, j_0) &= P(Y_{-f+1}, \ldots, Y_2, X_1 = j_1, X_2 = j_0) \\
&= P(Y_2 | Y_{-f+1}, \ldots, Y_1, X_1 = j_1, X_2 = j_0) \\
&\qquad \cdot P(X_2 = j_0 | Y_{-f+1}, \ldots, Y_1, X_1 = j_1) \\
&\qquad\quad \cdot P(Y_{-f+1}, \ldots, Y_1, X_1 = j_1) \\
&= P(Y_2 | Y_{-f+2}, \ldots, Y_1, X_2 = j_0) \\
&\qquad \cdot P(X_2 = j_0 | X_1 = j_1)\, P(Y_{-f+1}, \ldots, Y_1, X_1 = j_1) \\
&= c^{(j_0)}_{y_{-f+2}, \ldots, y_2}\, \pi_{j_0|j_1}\, \alpha_1(j_1)
\end{aligned}$$

For $t = 3, \ldots, \ell$, equation (4) is

$$
\begin{aligned}
\alpha_t(j_{t-1},\ldots,j_0) &= P(Y_{-f+1},\ldots,Y_t, X_1=j_{t-1},\ldots,X_t=j_0) \\
&= P(Y_t|Y_{-f+1},\ldots,Y_{t-1}, X_1=j_{t-1},\ldots,X_t=j_0) \\
&\quad \cdot P(X_t=j_0|Y_{-f+1},\ldots,Y_{t-1}, X_1=j_{t-1},\ldots,X_{t-1}=j_1) \\
&\quad\quad \cdot P(Y_{-f+1},\ldots,Y_{t-1}, X_1=j_{t-1},\ldots,X_{t-1}=j_1) \\
&= P(Y_t|Y_{t-f},\ldots,Y_{t-1}, X_t=j_0) \\
&\quad \cdot P(X_t=j_0|X_1=j_{t-1},\ldots,X_{t-1}=j_1) \\
&\quad\quad \cdot P(Y_{-f+1},\ldots,Y_{t-1}, X_1=j_{t-1},\ldots,X_{t-1}=j_1) \\
&= c^{(j_0)}_{y_{t-f},\ldots,y_t}\, \pi_{j_0|j_{t-1}\ldots j_1}\, \alpha_{t-1}(j_{t-1},\ldots,j_1)
\end{aligned}
$$

Finally, for $t=\ell+1,\ldots,T$, equation (5) is

$$
\begin{aligned}
&\alpha_t(j_{\ell-1},\ldots,j_0) \\
&= P(Y_{-f+1},\ldots,Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0) \\
&= P(Y_t|Y_{-f+1},\ldots,Y_{t-1}, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0) \\
&\quad \cdot P(Y_{-f+1},\ldots,Y_{t-1}, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0) \\
&= P(Y_t|Y_{t-f},\ldots,Y_{t-1}, X_t=j_0) \\
&\quad \cdot \sum_{j_\ell=1}^{M} P(Y_{-f+1},\ldots,Y_{t-1}, X_{t-\ell}=j_\ell,\ldots,X_t=j_0) \\
&= c^{(j_0)}_{y_{t-f},\ldots,y_t} \sum_{j_\ell=1}^{M} P(X_t=j_0|Y_{-f+1},\ldots,Y_{t-1}, X_{t-\ell}=j_\ell,\ldots,X_{t-1}=j_1) \\
&\quad \cdot P(Y_{-f+1},\ldots,Y_{t-1}, X_{t-\ell}=j_\ell,\ldots,X_{t-1}=j_1) \\
&= c^{(j_0)}_{y_{t-f},\ldots,y_t} \sum_{j_\ell=1}^{M} P(X_t=j_0|X_{t-\ell}=j_\ell,\ldots,X_{t-1}=j_1) \\
&\quad \cdot P(Y_{-f+1},\ldots,Y_{t-1}, X_{t-\ell}=j_\ell,\ldots,X_{t-1}=j_1) \\
&= c^{(j_0)}_{y_{t-f},\ldots,y_t} \sum_{j_\ell=1}^{M} a_{j_\ell,\ldots,j_0} \alpha_{t-1}(j_\ell,\ldots,j_1)
\end{aligned}
$$

The backward terms β_t are obtained as follows. For $t=\ell,\ldots,T-1$, equation (7) is

$$
\begin{aligned}
&\beta_t(j_{\ell-1},\ldots,j_0) \\
&= P(Y_{t+1},\ldots,Y_T|Y_{t-f},\ldots,Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{P(Y_{t-f},\ldots,Y_T, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0)}{P(Y_{t-f},\ldots,Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0)} \\
&= \frac{1}{P(Y_{t-f},\ldots,Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0)} \\
&\quad \cdot \sum_{j=1}^{M} P(Y_{t-f},\ldots,Y_T, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0, X_{t+1}=j) \\
&= \frac{1}{P(Y_{t-f},\ldots,Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0)} \\
&\quad \cdot \sum_{j=1}^{M} P(Y_{t-f},\ldots Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0) \\
&\quad\quad \cdot P(X_{t+1}=j|Y_{t-f},\ldots,Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0) \\
&\quad\quad\quad \cdot P(Y_{t+1}|Y_{t-f},\ldots,Y_t, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0, X_{t+1}=j) \\
&\quad\quad\quad\quad \cdot P(Y_{t+2},\ldots,Y_T|Y_{t-f},\ldots,Y_{t+1}, X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0, X_{t+1}=j) \\
&= \sum_{j=1}^{M} P(X_{t+1}=j|X_{t-\ell+1}=j_{\ell-1},\ldots,X_t=j_0) \\
&\quad \cdot P(Y_{t+1}|Y_{t-f+1},\ldots,Y_t, X_{t+1}=j) \\
&\quad\quad \cdot P(Y_{t+2},\ldots,Y_T|Y_{t-f+1},\ldots,Y_{t+1}, X_{t-\ell+2}=j_{\ell-2},\ldots,X_t=j_0, X_{t+1}=j) \\
&= \sum_{j=1}^{M} a_{j_{\ell-1},\ldots,j_0,j}\, c^{(j)}_{y_{t-f+1}\ldots y_{t+1}}\, \beta_{t+1}(j_{l-2},\ldots,j_0,j)
\end{aligned}
$$

and for $t = 1,\ldots,\ell-1$, equation (8) is

$$
\begin{aligned}
&\beta_t(j_{t-1},\ldots,j_0) \\
&\quad = P(Y_{t+1},\ldots,Y_T|Y_{t-f+1},\ldots,Y_t, X_1=j_{t-1},\ldots,X_t=j_0) \\
&\quad = \sum_{j=1}^{M} P(X_{t+1}=j|X_1=j_{t-1},\ldots,X_t=j_0) \\
&\quad\quad \cdot P(Y_{t+1}|Y_{t-f+1},\ldots,Y_t, X_{t+1}=j) \\
&\quad\quad\quad \cdot P(Y_{t+2},\ldots,Y_T|Y_{t-f+1},\ldots,Y_{t+1}, X_1=j_{t-1},\ldots,X_t=j_0, X_{t+1}=j) \\
&\quad = \sum_{j=1}^{M} \pi_{j|j_{t-1},\ldots,j_0}\, c^{(j)}_{y_{t-f+1}\ldots y_{t+1}}\, \beta_{t+1}(j_{t-1},\ldots,j_0,j)
\end{aligned}
$$

A.2 Estimation of π, A and C

For $t = \ell,\ldots,T-1$, equation (11) is obtained as

$$
\epsilon_t(j_{\ell-1},\ldots,j_0,j)
$$

$$
\begin{aligned}
&= P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0, X_{t+1} = j | Y_{-f+1}, \ldots, Y_T) \\
&= \frac{1}{P(Y_{-f+1}, \ldots, Y_T)} P(Y_{-f+1}, \ldots, Y_T, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0, X_{t+1} = j) \\
&= \frac{1}{P(Y_{-f+1}, \ldots, Y_T)} P(Y_{-f+1}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0) \\
&\quad \cdot P(X_{t+1} = j | Y_{-f+1}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0) \\
&\quad \cdot P(Y_{t+1} | Y_{-f+1}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, X_t = j_0, X_{t+1} = j) \\
&\quad \cdot P(Y_{t+2}, \ldots, Y_T | Y_{-f+1}, \ldots, Y_{t+1}, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0, X_{t+1} = j) \\
&= \frac{1}{P(Y_{-f+1}, \ldots, Y_T)} P(Y_{-f+1}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0) \\
&\quad \cdot P(X_{t+1} = j | X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0) \\
&\quad \cdot P(Y_{t+1} | Y_{t-f+1}, \ldots, Y_t, X_{t+1} = j) \\
&\quad \cdot P(Y_{t+2}, \ldots, Y_T | Y_{t-f+2}, \ldots, Y_{t+1}, X_{t-\ell+2} = j_{\ell-2}, \ldots, X_t = j_0, X_{t+1} = j) \\
&= \frac{\alpha_t(j_{\ell-1}, \ldots, j_0)\, a_{j_{\ell-1}, \ldots, j_0, j}\, c^{(j)}_{y_{t-f+1}, \ldots, y_{t+1}}\, \beta_{t+1}(j_{\ell-2}, \ldots, j_0, j)}{L(Y_{-f+1}, \ldots, Y_T)}
\end{aligned}
$$

For $t = \ell, \ldots, T$, equation (12) is obtained as

$$
\begin{aligned}
&\gamma_t(j_{\ell-1}, \ldots, j_0) \\
&= P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0 | Y_{-f+1}, \ldots, Y_T) \\
&= \frac{P(Y_{-f+1}, \ldots, Y_T, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0)}{P(Y_{-f+1}, \ldots, Y_T)} \\
&= \frac{1}{P(Y_{-f+1}, \ldots, Y_T)} P(Y_{-f+1}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0) \\
&\quad \cdot P(Y_{t+1}, \ldots, Y_T | Y_{-f+1}, \ldots, Y_t, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0) \\
&= \frac{\alpha_t(j_{\ell-1}, \ldots, j_0)\, \beta_t(j_{\ell-1}, \ldots, j_0)}{L(Y_{-f+1}, \ldots, Y_T)}
\end{aligned}
$$

For $t = 2, \ldots, \ell$, equation (13) is

$$
\begin{aligned}
\hat{\pi}_{j_0 | j_{t-1}, \ldots, j_1} &= P(X_t = j_0 | Y_{-f+1}, \ldots, Y_T, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1) \\
&= \frac{P(Y_{-f+1}, \ldots, Y_T, X_1 = j_{t-1}, \ldots, X_t = j_0)}{P(Y_{-f+1}, \ldots, Y_T, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1)} \\
&= \frac{\gamma_t(j_{t-1}, \ldots, j_0)}{\gamma_{t-1}(j_{t-1}, \ldots, j_1)}
\end{aligned}
$$

Equation (14) is computed as

$$
\hat{a}_{j_{\ell-1} \ldots j_0 j} = \sum_{t=\ell}^{T-1} P(X_{t+1} = j | X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0, Y_{-f+1}, \ldots, Y_T)
$$

$$
\begin{aligned}
&= \frac{\sum_{t=\ell}^{T-1} P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_{t+1} = j, Y_{-f+1}, \ldots, Y_T)}{\sum_{t=\ell}^{T-1} P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0, Y_{-f+1}, \ldots, Y_T)} \\
&= \frac{\sum_{t=\ell}^{T-1} P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_{t+1} = j | Y_{-f+1}, \ldots, Y_T)\, P(Y_{-f+1}, \ldots, Y_T)}{\sum_{t=\ell}^{T-1} P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0 | Y_{-f+1}, \ldots, Y_T)\, P(Y_{-f+1}, \ldots, Y_T)} \\
&= \frac{\sum_{t=\ell}^{T-1} P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_{t+1} = j | Y_{-f+1}, \ldots, Y_T)}{\sum_{t=\ell}^{T-1} P(X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0 | Y_{-f+1}, \ldots, Y_T)} \\
&= \frac{\sum_{t=\ell}^{T-1} \epsilon_t(j_{\ell-1}, \ldots, j_0, j)}{\sum_{t=\ell}^{T-1} \gamma_t(j_{\ell-1}, \ldots, j_0)}
\end{aligned}
$$

and equation (15) is

$$
\begin{aligned}
\hat{c}^{(j_0)}_{i_f \ldots i_0} &= P(Y_t = i_0 | Y_{-f+1}, \ldots, Y_{t-f} = i_f, \ldots, Y_{t-1} = i_1, Y_{t+1}, \ldots, Y_T, X_t = j_0) \\
&= \frac{P(Y_{-f+1}, \ldots, Y_{t-f} = i_f, \ldots, Y_{t-1} = i_1, Y_t = i_0, Y_{t+1}, \ldots, Y_T, X_t = j_0)}{P(Y_{-f+1}, \ldots, Y_{t-f} = i_f, \ldots, Y_{t-1} = i_1, Y_{t+1}, \ldots, Y_T, X_t = j_0)} \\
&= \frac{\sum_{j_{\ell-1}=1}^{M} \ldots \sum_{j_1=1}^{M} P(Y_{-f+1}, \ldots, Y_T, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0)}{\sum_{j_{\ell-1}=1}^{M} \ldots \sum_{j_1=1}^{M} P(Y_{-f+1}, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_T, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0)} \\
&= \frac{\sum_{j_{\ell-1}=1}^{M} \ldots \sum_{j_1=1}^{M} P(Y_{-f+1}, \ldots, Y_T, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0)}{\sum_{j_{\ell-1}=1}^{M} \ldots \sum_{j_1=1}^{M} \sum_{y_t=1}^{K} P(Y_{-f+1}, \ldots, Y_t = y_t, \ldots, Y_T, X_{t-\ell+1} = j_{\ell-1}, \ldots, X_t = j_0)}
\end{aligned}
$$

$$= \frac{\sum_{\substack{t=1 \\ Y_{t-f}=i_f \ldots Y_t=i_0}}^{T} \sum_{j_{\ell-1}=1}^{M} \cdots \sum_{j_1=1}^{M} \gamma_t(j_{\ell-1}, \ldots, j_0)}{\sum_{\substack{t=1 \\ Y_{t-f}=i_f \ldots Y_{t-1}=i_1}}^{T} \sum_{j_{\ell-1}=1}^{M} \cdots \sum_{j_1=1}^{M} \gamma_t(j_{\ell-1}, \ldots, j_0)}$$

A.3 Optimal sequence of hidden states

The quantities δ_t are obtained as follows. For $t = 1$ and $j_0 = 1, \ldots, M$, equation (16) is

$$\begin{aligned}
\delta_1(j_0) &= P(Y_{-f+1}, \ldots, Y_1, X_1 = j_0) \\
&= P(Y_{-f+1}, \ldots, Y_0, X_1 = j_0)\, P(Y_1 | Y_{-f+1}, \ldots, Y_0, X_1 = j_0) \\
&= P(X_1 = j_0)\, P(Y_1 | Y_{-f+1}, \ldots, Y_0, X_1 = j_0) \\
&= \pi_{j_0}\, c^{(j_0)}_{y_{-f+1}, \ldots, y_1}
\end{aligned}$$

For $t = 2, \ldots, \ell$ and $j_{t-1}, \ldots, j_0 = 1, \ldots, M$, equation (17) is

$$\begin{aligned}
\delta_t(j_{t-1}, \ldots, j_0) &= \max_{j_{t-1}, \ldots, j_1} P(Y_{-f+1}, \ldots, Y_t, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1, X_t = j_0) \\
&= \max_{j_{t-1}, \ldots, j_1} P(Y_{-f+1}, \ldots, Y_{t-1}, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1) \\
&\qquad \cdot P(X_t = j_0 | Y_{-f+1}, \ldots, Y_{t-1}, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1) \\
&\qquad\quad \cdot P(Y_t | Y_{-f+1}, \ldots, Y_{t-1}, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1, X_t = j_0) \\
&= \max_{j_{t-1}, \ldots, j_1} P(Y_{-f+1}, \ldots, Y_{t-1}, X_1 = j_{t-1}, \ldots, X_{t-1} = j_1) \\
&\qquad \cdot P(X_t = j_0 | X_1 = j_{t-1}, \ldots, X_{t-1} = j_1) \\
&\qquad\quad \cdot P(Y_t | Y_{t-f}, \ldots Y_{t-1}, X_t = j_0) \\
&= \left[\max_{j_{t-1}, \ldots, j_1} \delta_{t-1}(j_{t-1}, \ldots, j_1)\, \pi_{j_0 | j_1, \ldots, j_{t-1}} \right] c^{(j_0)}_{y_{t-f}, \ldots, y_t}
\end{aligned}$$

and for $t = \ell + 1, \ldots, T$ and $j_{\ell-1}, \ldots, j_0 = 1, \ldots, M$, equation (18) is

$$\begin{aligned}
\delta_t(j_{\ell-1}, \ldots, j_0) &= \max_{j_\ell, \ldots, j_1} P(Y_{-f+1}, \ldots, Y_t, X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1, X_t = j_0) \\
&= \max_{j_\ell, \ldots, j_1} P(Y_{-f+1}, \ldots, Y_{t-1}, X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1) \\
&\qquad \cdot P(X_t = j_0 | Y_{-f+1}, \ldots, Y_{t-1}, X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1) \\
&\qquad\quad \cdot P(Y_t | Y_{-f+1}, \ldots, Y_{t-1}, X_{t-\ell} = j_\ell, \ldots, X_{t-1} = j_1, X_t = j_0)
\end{aligned}$$

$$
\begin{aligned}
= \quad & \max_{j_\ell,\ldots,j_1} P(Y_{-f+1},\ldots,Y_{t-1}, X_{t-\ell}=j_\ell,\ldots,X_{t-1}=j_1) \\
& \quad \cdot P(X_t = j_0 | X_{t-\ell}=j_\ell,\ldots,X_{t-1}=j_1) \\
& \quad\quad \cdot P(Y_t | Y_{t-f},\ldots Y_{t-1}, X_t = j_0) \\
= \quad & \left[\max_{j_\ell,\ldots,j_1} \delta_{t-1}(j_\ell,\ldots,j_1)\, a_{j_\ell,\ldots,j_1,j_0}\right] c^{(j_0)}_{y_{t-f},\ldots,y_t}
\end{aligned}
$$