The Double Chain Markov Model

André Berchtold[1]

Technical Report no. 348
Department of Statistics
University of Washington
Seattle, WA 98195-4322.

February 1999

[1]Email: berchtol@stat.washington.edu, Andre.Berchtold@themes.unige.ch

Abstract

Among the class of discrete time Markovian processes, two models are widely used, the Markov chain and the Hidden Markov Model. A major difference between these two models lies in the relation between successive outputs of the observed variable. In a visible Markov chain, these are directly correlated while in hidden models they are not. However, in some situations it is possible to observe both a hidden Markov chain and a direct relation between successive observed outputs. Unfortunately, the use of either a visible or a hidden model implies the suppression of one of these hypothesis.

This paper presents a Markovian model called the Double Chain Markov Model which takes into account the main features of both visible and hidden models. Its main purpose is the modeling of non-homogeneous time-series. It is very flexible and can be estimated with traditional methods. The model is applied on a sequence of wind speeds and it appears to model data more successfully than both the usual Markov chains and hidden Markov models.

Keywords: Double Chain Markov Model (DCMM), Hidden Markov Model (HMM), Markov Chain (MC), Forward-Backward Algorithm, Baum-Welsh Algorithm, Viterbi Algorithm.

Contents

List of Tables

List of Figures

1 Introduction

A *Markov Chain* (MC) is a statistical model used to represent transitions between successive outputs of a discrete time random variable X_t (Dynkin, 1965, Kemeny & Snell, 1976, Kijima, 1997). This is an entirely visible process since each observed output is exactly identified with one state of the process. Althought it is widely used, this model can not handle all situations. In certain domains, such as speech recognition, there is no perfect identification between the state of the chain at time t and the corresponding output. At each time, the state of the chain is unknown and we observe the output of another variable Y_t, whose distribution depends on the state of the model. This process is called a *Hidden Markov Model* (HMM) (Rabiner, 1989, Elliott et al., 1995, MacDonald & Zucchini, 1997).

A major difference between these two Markov models lies in the relation between successive observed outputs. In the Markov chain, the output at time t depends directly on the output at time $t-1$ (see Figure 1). In the HMM, the outputs are conditionaly independent (Figure 2).

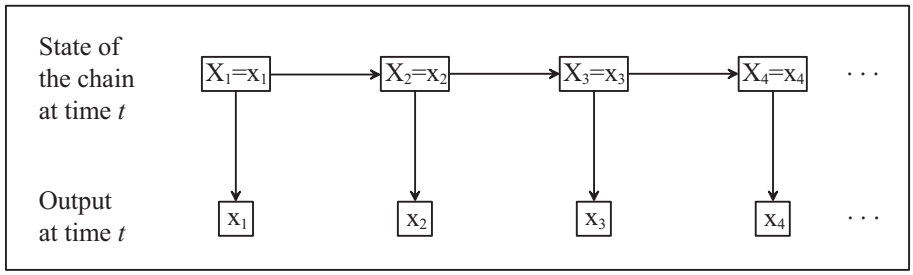

Figure 1 Markov Chain

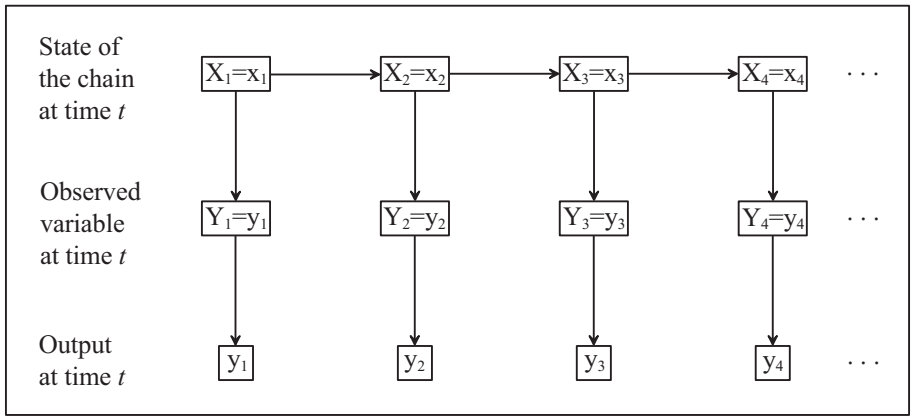

Figure 2 Hidden Markov Model

The conditional independence between outputs of a HMM is not always justified. It is entirely possible to observe processes governed by a hidden Markov chain in which successive observations are directly correlated. The speed of the wind (section 5.2) is an example. The *Double Chain Markov Model* (DCMM) presented in this paper combines characteristics of both visible and hidden models. It is called *double* since it can be viewed as the superposition of two Markov chains, a hidden chain governing the relation between states of a non-observable variable, and a visible chain governing the relation between outputs of an observed variable (Figure 3). Advantages of both Markov models are conserved: the system is driven by an unobserved process, but the successive outputs are directly correlated. Since the value of Y_1 depends on the past, we consider an initial output value at time 0 with no corresponding hidden state.

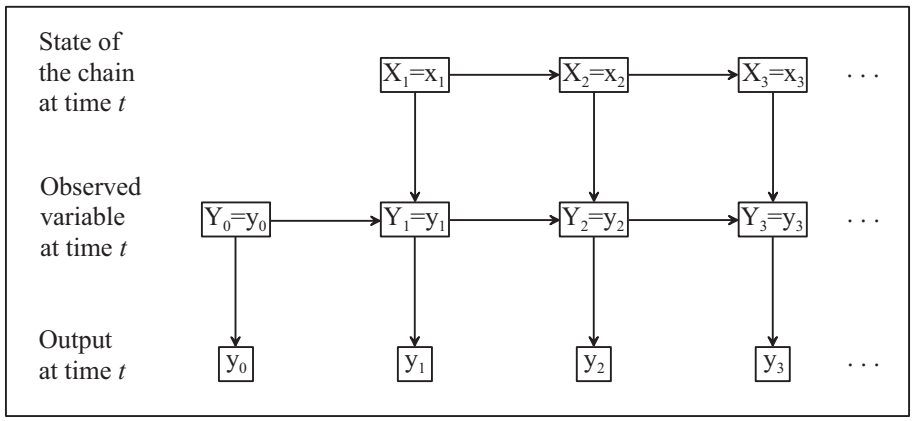

Figure 3 Double Chain Markov Model

The Double Chain Markov Model is designed for the modeling of non-homogeneous time-series. If a time-series can be decomposed into a finite set of transition matrices (Berchtold, 1998) the DCMM can be used to control the transition process between these matrices. Section 5 provides results of a simulation and a practical example.

The idea to include a direct relation between outputs into an Hidden Markov Model is not new. The first approach was to combine the HMM with an autoregressive model (Poritz, 1982, 1988, Kenny et al., 1990). Then, Wellekens (1987) and Paliwal (1993) presented a model similar to our DCMM, the former in the context of continuous HMM and the latter in the discrete case. Nevertheless, the present article has at least three main differences with the previous publications. Firstly, we provide a complete derivation of the Forward-Backward, Baum-Welch and Viterbi algorithms, while Paliwal gave only the main equation of the Forward algorithm, without justification. We give also computable versions of these

algorithms. Secondly, we interpret the relation between outputs as a non-stationnary Markov chain. Finally, Paliwal considered applications in speech recognition only, while we show that this model can be useful in other domains, such as in meteorology.

2 Model

The Double Chain Markov Model in discrete time combines two random variables: X_t whose state at time t is unknown for an observer external to the process, and Y_t which is observable. The model 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 hidden state, π=$\{\pi_1, \ldots, \pi_M\}$.
- A transition matrix between hidden states, $A = \{a_{ij}\}$, $i, j \in \mathcal{S}(X)$.
- A set of transition matrices between successive outputs of Y_t given a particular state of X_t, $C = \{c_{ijs}\}$, $i, j \in \mathcal{S}(Y)$, $s \in \mathcal{S}(X)$. C can also be written in a more convenient way as $C = \{C^{(s)}\}$ with $C^{(s)} = [c_{ij}^{(s)}]$.

A DCMM μ is then defined as $\mu = \{\pi, A, C\}$.

The Double Chain Markov Model is a generalization of both Markov chains and Hidden Markov Models. When there is only one hidden state ($M = 1$), the DCMM reduces to an homogeneous Markov chain with transition matrix $C^{(1)}$. On the other hand, when there is $M > 1$ hidden states but each matrix $C^{(s)}$ has identical rows, the model reduces to a HMM.

3 Estimation

We have to consider three different estimation problems:

1. The estimation of the likelihood of a sequence of observations $y_0, \ldots, y_T$ given a model.
2. This estimation of parameters π, A and C given a sequence of observations.
3. The estimation of the optimal sequence of hidden states given a model and a sequence of outputs.

These three problems are similar to the problems occuring in HMM theory and we will show that they can be solved using similar methods. The first problem is solved using a forward iterative algorithm. The estimation of the parameters is achieved with an Expectation-Maximization (EM) algorithm, and the optimal sequence of states is obtained through the Viterbi algorithm.

Remark: In this section, we give only the resulting formulas for each algorithm. The complete derivation is provided in appendix A. Moreover, as presented here, the three algorithms can lead to numerical problems since they involve the calculation of infinitesimal values. A good solution is to normalize the intermediary results at each step of the calculation. The practical implementation of this method is discussed in appendix B.

3.1 Likelihood of the observed output sequence

We want to calculate the likelihood of the observed sequence $y_0, \ldots, y_T$ given the model μ:

$$L = P(Y_0 = y_0, \ldots, Y_T = y_T | \mu) \tag{1}$$

This problem can be solved through an iterative procedure similar to the Forward procedure developped by Rabiner (1989) for the estimation of the HMM. We define

$$\alpha_t(j) = P(Y_0, \ldots, Y_t, X_t = j | \mu) \tag{2}$$

Remark: For readability, we write Y_t for $Y_t = y_t$. Moreover, we will not further indicate that the model μ is given in the calculation.

For $t = 1$ equation (2) becomes

$$\alpha_1(j) \;=\; c^{(j)}_{y_0 y_1}\, \pi_j \tag{3}$$

and in the general case, for $t = 2, \ldots, T$,

$$\alpha_t(j) \;=\; c^{(j)}_{y_{t-1} y_t} \sum_{i=1}^{M} a_{ij}\, \alpha_{t-1}(i) \tag{4}$$

The likelihood of the entire sequence of observations is obtained by summing $\alpha_T(j)$ over j:

$$
\begin{aligned}
L(Y_0, \ldots, Y_T) &= \sum_{j=1}^{M} P(Y_0, \ldots, Y_T, X_T = j) \\
&= \sum_{j=1}^{M} \alpha_T(j) \qquad (5)
\end{aligned}
$$

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 the parameters of the model. Let

$$
\beta_t(i) = P(Y_{t+1}, \ldots, Y_T | Y_t, X_t = i, \mu) \qquad (6)
$$

For $t = T$, we obtain

$$
\beta_T(i) = 1 \qquad (7)
$$

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

$$
\beta_t(i) = \sum_{j=1}^{M} a_{ij}\, c^{(j)}_{y_t y_{t+1}}\, \beta_{t+1}(j) \qquad (8)
$$

With this result, the likelihood can be rewritten as

$$
L(Y_0, \ldots, Y_T) = \sum_{i=1}^{M} \alpha_t(i)\, \beta_t(i) \quad , \quad t = 1, \ldots, T \qquad (9)
$$

Equation (5) corresponds to $t = T$.

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. We use an EM algorithm known in the speech recognition literature as the Baum-Welch algorithm. First, we define the joint probability of two successive hidden states. For $t = 1, \ldots, T-1$,

$$\begin{aligned}\epsilon_t(i,j) &= P(X_t = i, X_{t+1} = j | Y_0, \ldots, Y_T) \\ &= \frac{\alpha_t(i)\, a_{ij}\, c^{(j)}_{y_t y_{t+1}}\, \beta_{t+1}(j)}{L(Y_0, \ldots, Y_T)}\end{aligned} \tag{10}$$

Then we define the marginal distribution of the hidden states. For $t = 1, \ldots, T$,

$$\begin{aligned}\gamma_t(i) &= P(X_t = i | Y_0, \ldots, Y_T) \\ &= \frac{\alpha_t(i)\, \beta_t(i)}{L(Y_0, \ldots, Y_T)}\end{aligned} \tag{11}$$

The following relation holds for $t = 1, \ldots, T-1$:

$$\gamma_t(i) = \sum_{j=1}^{M} \epsilon_t(i,j) \tag{12}$$

Using ϵ_t and γ_t we can write the reestimation formulas for π, A and C as follows:

$$\begin{aligned}\hat{\pi}_i &= P(X_1 = i | Y_0, \ldots, Y_T) \\ &= \gamma_1(i)\end{aligned} \tag{13}$$

$$\begin{aligned}\hat{a}_{ij} &= \sum_{t=1}^{T-1} P(X_{t+1} = j | X_t = i, Y_0, \ldots, Y_T) \\ &= \frac{\sum_{t=1}^{T-1} \epsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}\end{aligned} \tag{14}$$

$$\begin{aligned}\hat{c}^{(i)}_{hk} &= P(Y_t = k | Y_0, \ldots, Y_{t-1} = h, \ldots, Y_T, X_t = i) \\ &= \frac{\sum_{\substack{t=1 \\ Y_t = k \\ Y_{t-1} = h}}^{T} \gamma_t(i)}{\sum_{\substack{t=1 \\ Y_{t-1} = h}}^{T} \gamma_t(i)}\end{aligned} \tag{15}$$

In practice, the estimation of the model is achieved using iteratively the Forward-Backward procedure and the reestimation formulas for π, A and C. Since we cannot insure that this procedure converges to the global maximum of the likelihood rather than to a local maximum, the choice of starting values is critical.

3.3 Optimal sequence of hidden states

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

$$P(X_1, \ldots, X_T | Y_0, \ldots, Y_T) \tag{16}$$

or the joint probability

$$P(X_1, \ldots, X_T, Y_0, \ldots, Y_T) \tag{17}$$

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

$$\begin{aligned} \delta_1(j) &= P(Y_0, Y_1, X_1 = j) \\ &= \pi_j \, c^{(j)}_{y_0 y_1} \end{aligned} \tag{18}$$

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

$$\begin{aligned} \delta_t(j) &= \max_{i_1, \ldots, i_{t-1}} P(Y_0, \ldots, Y_t, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = j) \\ &= \left[\max_{i_1, \ldots, i_{t-1}} \delta_{t-1}(i_{t-1}) \, a_{i_{t-1} j} \right] c^{(j)}_{y_{t-1} y_t} \end{aligned} \tag{19}$$

The optimal hidden state at time T is then determined as

$$\hat{x}_T = \arg \max_{j=1,\ldots,M} \delta_T(j) \tag{20}$$

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

$$\hat{x}_t = \arg \max_{j=1,\ldots,M} \delta_{t+1}(j)\, a_{j\hat{x}_{t+1}} \tag{21}$$

Finally, the joint probability of the sequence of hidden states and the sequence of observed outputs is equal to

$$P(X_1, \ldots, X_T, Y_0, \ldots, Y_T) = \max_{j=1,\ldots,M} \delta_T(j) \tag{22}$$

4 Simultaneous data

In many situations, the data consists of several time-series. For instance, we can observe the daily temperature in a given place during the month of February for 20 years. We have then 20 independent sequences of data, each of which can be of a different length. We note N the number of sequences, S_n the n^{th} sequence, T_n its length (without counting the observation at time 0), X_t^n its t^{th} hidden state and Y_t^n its t^{th} observed output. We want to model the whole set of data with only one DCMM μ, so we have only one distribution π, one transition matrix A and one set of transition matrices C to estimate.

Since all sequences are considered independent, the most of the calculation can take place separately upon each sequence. The likelihood $L(S_n)$ of the n^{th} sequence is obtain by computing the Forward procedure and applying equation (5). The global likelihood of all data is then equal to

$$L(S_1, \ldots, S_N) = \prod_{n=1}^{N} L(S_n) \tag{23}$$

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

$$\begin{aligned} \hat{\pi}_i &= P(X_1^n = i | Y_0^n, \ldots, Y_{T_n}^n) \quad , \quad \forall n \\ &= \frac{\sum_{n=1}^{N} \gamma_1^n(i)}{N} \end{aligned} \tag{24}$$

$$\begin{aligned}\hat{a}_{ij} &= \sum_{t=1}^{T_n-1} P(X_{t+1}^n = j | X_t^n = i, Y_0^n, \ldots, Y_{T_n}^n) \quad , \quad \forall n \\ &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n-1} \epsilon_t^n(i,j)}{\sum_{n=1}^{N} \sum_{t=1}^{T_n-1} \gamma_t^n(i)} \end{aligned} \tag{25}$$

$$\begin{aligned}\hat{c}_{hk}^{(i)} &= P(Y_t^n = k | Y_0, \ldots, Y_{t-1}^n = h, \ldots, Y_{T_n}^n, X_t^n = i) \quad , \quad \forall n \\ &= \frac{\sum_{n=1}^{N} \sum_{\substack{t=1 \\ Y_t^n = k \\ Y_{t-1}^n = h}}^{T_n} \gamma_t^n(i)}{\sum_{n=1}^{N} \sum_{\substack{t=1 \\ Y_{t-1}^n = h}}^{T_n} \gamma_t^n(i)} \end{aligned} \tag{26}$$

Finally, 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 obtained as

$$P(S(X_1), S(Y_1); \ldots; S(X_N), S(Y_N)) = \prod_{n=1}^{N} P(S(X_n), S(Y_n)) \tag{27}$$

on account of the independence of each sequence.

5 Applications

In this section we present a set of simulations and a real application using the Double Chain Markov Model. All comparisons between models were carried out using the Bayes Information Criterion (BIC). This criterion is defined as

$$BIC = -2LL + p \log(n) \tag{28}$$

where LL is the log-likelihood of the model, p its number of independent parameters and n the number of data. The model achieving the lowest BIC is chosen. See Katz (1981) for a discussion of the use of this criterion in the context of Markov chains.

Remark: According to the convention established in Bishop et al. (1975), we did not count the parameters that were equal to zero.

5.1 Simulations

Our first experiment is to verify that the Double Chain Markov Model can represent non-homogeneous time-series. We defined the following three states two outputs DCMM:

$$A = \begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 0.1 & 0.9 \end{pmatrix} \quad , \quad C^{(1)} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

$$C^{(2)} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \quad , \quad C^{(3)} = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix}$$

In the first state, the next output has a large probability (90%) to be identical as the previous one, and the third state is the opposite of the first. The second state represents independence. The probability to stay in the same hidden state the next period is always greather or equal to 80% and it is not possible to go in one step from state one to state three and vice versa. We used this model to generate 20 sequences of length 504 and we tried 11 different models upon each sequence: the independence model, Markov chains of order 1 to 4, HMM with 2, 3 and 4 hidden states and DCMM with 2, 3 and 4 hidden states. The first data of each sequence were dropped in order to have the same number of elements (500) in the log-likelihood of each model. We classified the models according to their BIC value and Table 1 summarizes the number of first, second and third rank obtained by each model.

According to Table 1, the best model is almost always a DCMM, but two hidden states are generally sufficient, which is somewhat surprising since the data were generated by a model with three hidden states. The explanation is found by looking at one of the estimated models (here, for the third simulation):

$$\widehat{A} = \begin{pmatrix} 0.9611 & 0.0389 \\ 0.0525 & 0.9475 \end{pmatrix}$$

$$\widehat{C}^{(1)} = \begin{pmatrix} 0.1631 & 0.8369 \\ 0.8453 & 0.1547 \end{pmatrix} \quad , \quad \widehat{C}^{(2)} = \begin{pmatrix} 0.7246 & 0.2754 \\ 0.1575 & 0.8425 \end{pmatrix}$$

Table 1. Classification of the models according to their BIC. 20 datasets were simulated. For 14 of these, the DCMM 2 was best, in 8 of them, the DCMM 3 was second-best, and so on.

Model	Rank		
	1	2	3
Independence	-	-	-
MC 1	-	-	-
MC 2	1	1	2
MC 3	-	4	5
MC 4	-	-	-
HMM 2	-	-	1
HMM 3	-	-	1
HMM 4	-	4	2
DCMM 2	14	1	-
DCMM 3	4	8	3
DCMM 4	1	2	6

The independence situation corresponding to the second hidden state of the original model does not appear explicitly. The two remaining states correspond roughly to the two others original states ($\widehat{C}^{(1)}$ corresponds to $C^{(3)}$ and $\widehat{C}^{(2)}$ to $C^{(1)}$) but the probabilities are less extreme since these two states must also take into account the data generated by $C^{(2)}$. What is happening is that the model concentrates upon the informative part of the data and does not add extra-parameters for the non-informative independence situation.

This example shows that not only can the Double Chain Markov Model successfully represent non-homogeneous time-series, but it can also do so better either than the homogeneous Markov chains or the hidden Markov models.

5.2 High and Low Wind Speeds

We consider a length 6574 time-series of the daily average wind speed at Roche's Point (South of Ireland) during the period 1961-1978. These data were previously analyzed in Haslett & Raftery (1989).

We are interested in the possible use of the wind power. More specifically, we want to model two extreme situations which can prevent a good exploitation of this power: days with exceptionnal low and high wind speed. Accordingly, we classified the data into three categories as given in Table 2. The wind speed is given in knots (1 knot = 0.5148 m/s).

Table 2 Classification of the wind speed data

Category	Wind speed	Number of days
low	< 5	494
normal	[5-20]	5437
high	> 20	643

The speed of the wind a given day is correlated with its speed the previous day, but the process is not stationnary and evolves throughout the year. Hence a DCMM is well suited to represent these data. We computed also several other models including the independence model, Markov chains of order 1 to 4, and Hidden Markov Models of orders 2 and 3. We tried also to model high-order Markov chains with the MTD model (Raftery, 1985, Raftery & Tavaré, 1994, Berchtold, 1996). Table 3 reports our results. For the purpose of comparison, we dropped the initial observations of the time-series in order to have the same number of elements (6570) in the log-likelihood of each model.

Table 3 Different modelings of the wind speed data

Model	Number of parameters	Log-likelihood	BIC
Independence	2	-3805.1	7627.9
MC 1	6	-3508.2	7069.1
MC 2	14	-3491.2	7105.4
MC 3	30	-3469.5	7202.8
MC 4	60	-3434.7	7396.7
MTD 2	5	-3499.7	7043.3
MTD 3	6	-3494.6	7042.0
MTD 4	7	-3490.1	7041.8
HMM 2	5	-3577.8	7199.5
HMM 3	9	-3476.1	7031.3
DCMM 2	12	-3448.2	**7001.9**
DCMM 3	15	-3445.9	7023.6

According to the BIC values of Table 3, the best model is the DCMM with two hidden states whose parameters are

$$\pi = \begin{pmatrix} 0 & 1 \end{pmatrix}$$

$$A = \begin{pmatrix} 0.9875 & 0.0125 \\ 0.0148 & 0.9852 \end{pmatrix}$$

$$C^{(1)} = \begin{pmatrix} 0.3550 & 0.6450 & 0 \\ 0.0805 & 0.8874 & 0.0321 \\ 0.0228 & 0.7721 & 0.2051 \end{pmatrix}$$

$$C^{(2)} = \begin{pmatrix} 0.1973 & 0.7846 & 0.0181 \\ 0.0361 & 0.8137 & 0.1502 \\ 0 & 0.6826 & 0.3174 \end{pmatrix}$$

We note first that the choice of a model with a first-order dependence between successive days is consistent with the other results of Table 3. Effectively, the best Markov chain is of order 1, and although the MTD model improves the results, the fourth order MTD cannot compete with the DCMM 2.

In the DCMM 2, each state represents mainly the second category of data (normal wind speed) but with a tendency to handle also either low or high speeds. The transition matrix $C^{(1)}$ corresponding to the first hidden state is used to represent low speeds. By comparison with the other matrix, the probability to go from a normal or a high speed to a low speed is always greater in $C^{(1)}$. The probabilty to stay from day to day in a situation of low speed (0.355) is also greater. Finally, it is impossible to go directly from a day with low speed to a day with high speed. Similar observations can be made on the transition matrix $C^{(2)}$ for the case of high wind speeds. The probability to stay in the same state from day to day is very high (0.9875 for the first state and 0.9852 for the second state). This indicates that the system is not often likely to switch from a situation of low speeds to a situation of high speeds and vice versa.

6 Developments

The Double Chain Markov Model as presented in this paper is well-defined and applicable in many different situations. However, it can be developed and improved in several ways. First, we considered only first-order transitions between both the hidden states and the outputs. To allow longer time dependencies, it is possible to replace the first-order transition matrices A and C by high-order matrices.

Even in the most simple case, the main limitation of the DCMM is its large number of parameters. The use of high-order transitions could only reinforce this problem. It is then necessary to consider a modeling of the transition matrices. The Mixture Transition Distribution (MTD) model (Raftery, 1985, Raftery & Tavaré, 1994, Berchtold, 1996) is a good solution for the modeling of high-order transitions. Moreover, it has already been used in the case of Hidden Markov Models (Schimert, 1992).

In the DCMM, the only phenomenon influencing the variable Y_t other than its past is a hidden process. In real situations, another visible process can influence Y_t too. This could be handle by the introduction of covariates modifying the transition matrices A and C.

7 Conclusion

The modeling of homogeneous discrete time-series has been studied for a long time with different types of models, either markovian or non-markovian. However, the markovian study of non-homogeneous time-series has not been treated extensively. In this paper we presented a model called the Double Chain Markov Model (DCMM), combining the Hidden Markov Model with a non-homogeneous Markov chain. The result is a very general markovian framework in which both HMM and traditional chains are particular cases. This model can be estimated using standard methods. Simulations show that the DCMM handles correctly non-homogeneous situations and proves to be better than both high-order Markov chains and Hidden Markov Models.

Acknowledgements

This work was supported by a grant from the Swiss National Science Foundation. I would like to thank Gilles Celeux, Chris Fraley, Alejandro Murua, Adrian Raftery and Gilbert Ritschard for their very helpful comments.

References

Berchtold, A. (1996) Modélisation autorégressive des chaînes de Markov : Utilisation d'une matrice différente pour chaque retard. *Revue de Statistique Appliquée*, Vol. XLIV (3), 5-25.

BERCHTOLD, A. (1998) Learning in Markov Chains. In *Apprentissage, des principes naturels aux méthodes artificielles*. Ritschard, Berchtold, Duc & Zighed Editors, HERMES, Paris.

BISHOP, Y. M. M., S. E. FIENBERG, P. W. HOLLAND (1975) *Discrete Multivariate Analysis*. MIT Press, Cambridge.

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.

KEMENY, J. G., J. L. SNELL, A. W. KNAPP (1976) *Denumerable 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.

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., TAVARÉ, S. (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.

A Derivation of the algorithms

In this appendix, we provide the complete derivation of the algorithms of section 3.

A.1 Likelihood of the observed output sequence

The two formulas for the calculation of α_t were obtained as follows. For $t = 1$ equation (2) becomes

$$\begin{aligned} \alpha_1(j) &= P(Y_0, Y_1, X_1 = j) \\ &= P(Y_1 | Y_0, X_1 = j)\, P(Y_0, X_1 = j) \\ &= c^{(j)}_{y_0 y_1}\, P(Y_0, X_1 = j) \end{aligned}$$

Since Y_0 and X_1 are independent and the value of Y_0 is known,

$$\begin{aligned} \alpha_1(j) &= c^{(j)}_{y_0 y_1}\, P(Y_0)\, P(X_1 = j) \\ &= c^{(j)}_{y_0 y_1}\, P(X_1 = j) \\ &= c^{(j)}_{y_0 y_1}\, \pi_j \end{aligned}$$

For $t > 1$, equation (4) becomes

$$\begin{aligned}
\alpha_t(j) &= P(Y_0, \ldots, Y_t, X_t = j) \\
&= P(Y_t | Y_0, \ldots, Y_{t-1}, X_t = j)\, P(Y_0, \ldots, Y_{t-1}, X_t = j) \\
&= P(Y_t | Y_{t-1}, X_t = j) \sum_{i=1}^{M} P(Y_0, \ldots, Y_{t-1}, X_{t-1} = i, X_t = j) \\
&= c^{(j)}_{y_{t-1} y_t} \sum_{i=1}^{M} P(X_t = j | Y_0, \ldots, Y_{t-1}, X_{t-1} = i) \\
&\qquad \cdot P(Y_0, \ldots, Y_{t-1}, X_{t-1} = i) \\
&= c^{(j)}_{y_{t-1} y_t} \sum_{i=1}^{M} P(X_t = j | X_{t-1} = i)\, P(Y_0, \ldots, Y_{t-1}, X_{t-1} = i) \\
&= c^{(j)}_{y_{t-1} y_t} \sum_{i=1}^{M} a_{ij}\, \alpha_{t-1}(i)
\end{aligned}$$

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

$$\begin{aligned}
\beta_t(i) &= P(Y_{t+1}, \ldots, Y_T | Y_t, X_t = i) \\
&= \frac{P(Y_t, \ldots, Y_T, X_t = i)}{P(Y_t, X_t = i)} \\
&= \frac{1}{P(Y_t, X_t = i)} \sum_{j=1}^{M} P(Y_t, \ldots, Y_T, X_t = i, X_{t+1} = j) \\
&= \frac{1}{P(Y_t, X_t = i)} \sum_{j=1}^{m} P(Y_t, X_t = i)\, P(X_{t+1} = j | Y_t, X_t = i) \\
&\qquad \cdot P(Y_{t+1} | Y_t, X_t = i, X_{t+1} = j) \\
&\qquad \cdot P(Y_{t+2}, \ldots,_T | Y_t, Y_{t+1}, X_t = i, X_{t+1} = j) \\
&= \sum_{j=1}^{M} P(X_{t+1} = j | X_t = i)\, P(Y_{t+1} | Y_t, X_{t+1} = j) \\
&\qquad \cdot P(Y_{t+2}, \ldots,_T | Y_t, Y_{t+1}, X_t = i, X_{t+1} = j) \\
&= \sum_{j=1}^{M} a_{ij}\, c^{(j)}_{y_t y_{t+1}}\, \beta_{t+1}(j)
\end{aligned}$$

A.2 Estimation of π, A and C

Equation (10) for the calculation of $\epsilon_t(i,j)$ is

$$\begin{aligned}
\epsilon_t(i,j) &= P(X_t = i, X_{t+1} = j | Y_0, \ldots, Y_T) \\
&= \frac{1}{P(Y_0, \ldots, Y_T)} P(Y_0, \ldots, Y_T, X_t = i, X_{t+1} = j) \\
&= \frac{1}{P(Y_0, \ldots, Y_T)} P(Y_0, \ldots, Y_t, X_t = i) \, P(X_{t+1} | Y_0, \ldots, Y_t, X_t = i) \\
&\qquad \cdot P(Y_{t+1} | Y_0, \ldots, Y_t, X_t = i, X_{t+1} = j) \\
&\qquad\quad \cdot P(Y_{t+2}, \ldots, Y_T | Y_0, \ldots, Y_{t+1}, X_t = i, X_{t+1} = j) \\
&= \frac{1}{P(Y_0, \ldots, Y_T)} P(Y_0, \ldots, Y_t, X_t = i) \, P(X_{t+1} | X_t = i) \\
&\qquad \cdot P(Y_{t+1} | Y_t, X_{t+1} = j) \, P(Y_{t+2}, \ldots, Y_T | Y_{t+1}, X_{t+1} = j) \\
&= \frac{\alpha_t(i) \, a_{ij} \, c^{(j)}_{y_t y_{t+1}} \, \beta_{t+1}(j)}{L(Y_0, \ldots, Y_T)}
\end{aligned}$$

and, for $\gamma_t(i)$, we have

$$\begin{aligned}
\gamma_t(i) &= P(X_t = i | Y_0, \ldots, Y_T) \\
&= \frac{P(Y_0, \ldots, Y_T, X_t = i)}{P(Y_0, \ldots, Y_T)} \\
&= \frac{P(Y_0, \ldots, Y_t, X_t = i) \, P(Y_{t+1}, \ldots, Y_T | Y_0, \ldots, Y_t, X_t = i)}{P(Y_0, \ldots, Y_T)} \\
&= \frac{P(Y_0, \ldots, Y_t, X_t = i) P(Y_{t+1}, \ldots, Y_T | Y_t, X_t = i)}{P(Y_0, \ldots, Y_T)} \\
&= \frac{\alpha_t(i) \, \beta_t(i)}{L(Y_0, \ldots, Y_T)}
\end{aligned}$$

Then

$$\begin{aligned}
\hat{a}_{ij} &= \sum_{t=1}^{T-1} P(X_{t+1} = j | X_t = i, Y_0, \ldots, Y_T) \\
&= \frac{\sum_{t=1}^{T-1} P(X_t = i, X_{t+1} = j, Y_0, \ldots, Y_T)}{\sum_{t=1}^{T-1} P(X_t = i, Y_0, \ldots, Y_T)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{t=1}^{T-1} P(X_t = i, X_{t+1} = j | Y_0, \ldots, Y_T)\, P(Y_0, \ldots, Y_T)}{\sum_{t=1}^{T-1} P(X_t = i | Y_0, \ldots, Y_T)\, P(Y_0, \ldots, Y_T)} \\
&= \frac{\sum_{t=1}^{T-1} P(X_t = i, X_{t+1} = j | Y_0, \ldots, Y_T)}{\sum_{t=1}^{T-1} P(X_t = i | Y_0, \ldots, Y_T)} \\
&= \frac{\sum_{t=1}^{T-1} \epsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{c}_{hk}^{(i)} &= P(Y_t = k | Y_0, \ldots, Y_{t-2}, Y_{t-1} = h, Y_{t+1}, \ldots, Y_T, X_t = i) \\
&= \frac{P(Y_0, \ldots, Y_{t-2}, Y_{t-1} = h, Y_t = k, Y_{t+1}, \ldots, Y_T, X_t = i)}{P(Y_0, \ldots, Y_{t-2}, Y_{t-1} = h, Y_{t+1}, \ldots, Y_T, X_t = i)} \\
&= \frac{\sum_{\substack{t=1 \\ Y_t = k \\ Y_{t-1} = h}}^{T} \gamma_t(i)}{\sum_{\substack{t=1 \\ Y_{t-1} = h}}^{T} \gamma_t(i)}
\end{aligned}
$$

A.3 Optimal sequence of hidden states

Equation (18) was obtained as

$$
\begin{aligned}
\delta_1(j) &= P(Y_0, Y_1, X_1 = j) \\
&= P(Y_0, X_1 = j)\, P(Y_1 | Y_0, X_1 = j) \\
&= P(X_1 = j)\, P(Y_1 | Y_0, X_1 = j) \\
&= \pi_j\, c_{y_0 y_1}^{(j)}
\end{aligned}
$$

and, for $t = 2, \ldots, T$ and $j = 1, \ldots, M$,

$$\begin{aligned}
\delta_t(j) &= \max_{i_1,\ldots,i_{t-1}} P(Y_0, \ldots, Y_t, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = j) \\
&= \max_{i_1,\ldots,i_{t-1}} P(Y_0, \ldots, Y_{t-1}, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}) \\
&\qquad \cdot P(X_t = j | Y_0, \ldots, Y_{t-1}, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}) \\
&\qquad\quad \cdot P(Y_T | Y_0, \ldots, Y_{t-1}, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = j) \\
&= \max_{i_1,\ldots,i_{t-1}} P(Y_0, \ldots, Y_{t-1}, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}) \\
&\qquad \cdot P(X_t | X_{t-1} = i_{t-1})\, P(Y_T | Y_{t-1}, X_t = j) \\
&= \left[\max_{i_1,\ldots,i_{t-1}} \delta_{t-1}(i_{t-1})\, a_{i_{t-1}j} \right] c^{(j)}_{y_{t-1}y_t}
\end{aligned}$$

B Practical computation of the algorithms

We mentioned that some algorithms used in this article lead to numerical errors. We provide in this appendix a computable version of these algorithms. Additional information can be found in MacDonald & Zucchini (1997).

B.1 Computation of the forward procedure

The forward terms $\alpha_t(i)$ can easely take values too small to be handled by a computer. To avoid this problem, one solution is to normalize $\alpha_t(i)$ at each step t. Formally, we define for $t = 1$ and $i = 1, \ldots, M$

$$\tilde{\alpha}_1(i) = \frac{c^{(i)}_{y_0 y_1}\, \pi_i}{\bar{\alpha}_1}$$

where

$$\bar{\alpha}_1 = \frac{\displaystyle\sum_{j=1}^{M} c^{(i)}_{y_0 y_1}\, \pi_i}{M}$$

is the average value of the $\alpha_1(i)$. Then, for $t = 2, \ldots, T$,

$$\tilde{\alpha}_t(j) \quad = \quad \frac{c^{(j)}_{y_{t-1}y_t} \sum_{i=1}^{M} a_{ij}\, \tilde{\alpha}_{t-1}(i)}{\bar{\alpha}_t}$$

where

$$\bar{\alpha}_t = \frac{c^{(j)}_{y_{t-1}y_t} \sum_{i=1}^{M} a_{ij}\, \tilde{\alpha}_{t-1}(i)}{M}$$

and the log-likelihood of the sequence of observations is obtained as

$$\log\{L(Y_0,\ldots,Y_T)\} \quad = \quad \log\left\{\sum_{i=1}^{M} \tilde{\alpha}_T(i)\right\} + \sum_{t=1}^{T} \log(\bar{\alpha}_t)$$

B.2 Computation of the backward procedure

The computation of the backward procedure leads to the same kind of problems as the forward procedure and requires the same type of solution. For $t = T$ and $i = 1,\ldots,M$ we define

$$\tilde{\beta}_T(i) = 1$$

and $\bar{\beta}_T = 1$. Then, for $t = T-1,\ldots,1$, we have

$$\tilde{\beta}_t(i) \quad = \quad \frac{\sum_{j=1}^{M} c^{(j)}_{y_t y_{t+1}}\, \tilde{\beta}_{t+1}(j)\, a_{ij}}{\bar{\beta}_t}$$

where

$$\bar{\beta}_t = \frac{\sum_{j=1}^{M} c^{(j)}_{y_t y_{t+1}}\, \tilde{\beta}_{t+1}(j)\, a_{ij}}{M}$$

For $t = 1, \ldots, T$ the log-likelihood is obtained as

$$\log \{L(Y_0, \ldots, Y_T)\} = \log \left\{ \sum_{i=1}^{M} \tilde{\alpha}_t(i) \tilde{\beta}_t(i) \right\} + \sum_{j=1}^{t} \log(\bar{\alpha}_j) + \sum_{j=t}^{T} \log(\bar{\beta}_j)$$

B.3 Computation of ϵ_t and γ_t

The computation of ϵ_t and γ_t is achieved using the scaled versions $\tilde{\alpha}_t$ and $\tilde{\beta}_t$ of α_t and β_t. For $t = 1, \ldots, T-1$,

$$\begin{aligned} \epsilon_t(i,j) = \exp \Bigg\{ & \log(\tilde{\alpha}_t(i)) + \sum_{k=1}^{t} \log(\bar{\alpha}_t) + \log(a_{ij}) + \log(c^{(j)}_{y_t y_{t+1}}) \\ & + \log(\tilde{\beta}_t(i)) + \sum_{k=t}^{T-1} \log(\bar{\beta}_t) - \log \{L(Y_0, \ldots, Y_T)\} \Bigg\} \end{aligned}$$

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

$$\begin{aligned} \gamma_t(i) = \exp \Bigg\{ & \log(\tilde{\alpha}_t(i)) + \sum_{k=1}^{t} \log(\bar{\alpha}_t) + \log(\tilde{\beta}_t(i)) \\ & + \sum_{k=t}^{T} \log(\bar{\beta}_t) - \log \{L(Y_0, \ldots, Y_T)\} \Bigg\} \end{aligned}$$

For $t = 1, \ldots, T-1$, the relation (12) holds. The reestimation of π, A and C is obtained through formulas (13), (14) and (15).

B.4 Computation of the optimal sequence of hidden states

To avoid numerical errors, it is necessary to scale the quantity δ used in the Viterbi algorithm. For $t = 1$ and $j = 1, \ldots, M$ we define

$$\tilde{\delta}_1(j) = \frac{\pi_j \, c^{(j)}_{y_0 y_1}}{\bar{\delta}_1}$$

where

$$\bar{\delta}_1 = \frac{\sum_{i=1}^{M} \pi(i)\, c^{(j)}_{y_0 y_1}}{M}$$

For $t = 2, \ldots, T$ and $j = 1, \ldots, M$ we compute iteratively

$$\tilde{\delta}_t(j) \;=\; \frac{\left[\max_{i_1,\ldots,i_{t-1}} a_{i_{t-1}j}\, \delta_{t-1}(i_{t-1})\right] c^{(j)}_{y_{t-1} y_t}}{\bar{\delta}_t}$$

and

$$\bar{\delta}_t = \frac{\sum_{i=1}^{M} \left[\max_{i_1,\ldots,i_{t-1}} a_{i_{t-1}j}\, \delta_{t-1}(i_{t-1})\right] c^{(j)}_{y_{t-1} y_t}}{M}$$

The optimal hidden state at time T is then

$$\hat{x}_T = \arg \max_{j=1,\ldots,m} \tilde{\delta}_T(j)$$

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

$$\hat{x}_t = \arg \max_{j=1,\ldots,m} \tilde{\delta}_{t+1}(j)\, a_{j\hat{x}_{t+1}}$$

Finally, the joint probability of the sequence of hidden states and the sequence of observed outputs is equal to

$$P(X_1, \ldots, X_T, Y_0, \ldots, Y_T) = \left[\max_{j=1,\ldots,M} \tilde{\delta}_T(j)\right] \prod_{t=1}^{T} \bar{\delta}_t$$