Mai 2004 **Numéro 36**

Cahiers de l'IMA

Confidence Intervals for Markovian Models

André Berchtold

Institut de Mathématiques Appliquées
Faculté des S.S.P.
Université de Lausanne
BFSH 2
CH-1015 Lausanne

Confidence Intervals for Markovian Models

André Berchtold

University of Lausanne, Switzerland

Abstract. This paper introduces a new multinomial approach unifying the computation of confidence intervals for Markovian models. Starting from a method used for homogeneous Markov chains, we show that it can be applied on models incorporating a hidden component. We consider three models derived from the basic homogeneous Markov chain: the Mixture Transition Distribution (MTD) model, the Hidden Markov Model (HMM), and the Double Chain Markov Model (DCMM). Compared to existing methods, our proposal can be used on data sets of any size without requiring extensive computing.

Keywords: Confidence interval; Markov chain; Mixture transition distribution model; Hidden Markov model; Double chain Markov model

1. Introduction

Markovian models such as homogeneous Markov chains and Hidden Markov models (HMMs) are very useful for the analysis of time series. They are, for instance, a standard tool in speech recongnition and DNA analysis, and they become more and more popular in psychology and social sciences. When a very large literature exists on Markov chains, some aspects of these models have not been fully studied yet, including the construction of confidence intervals for transition probabilities. Even if this point is a crucial one for the evaluation of the modeling quality, many recent publications totally ignore this aspect. The purpose of this paper is to provide an easy to compute method for the evaluation of simultameous confidence intervals in Markovian models.

The statistical literature adresses the construction of confidence intervals for the parameters of many models. However, this issue was never prominent in the case of Markovian models. Two reasons can explain that fact: i) Markov chains were often used in a theoretical framework which did not require real data; ii) even if models such as HMMs are of great use today, they are often applied in fields where the number of available data is very large, suppressing the need to worry about sample size and confidence intervals. In the case of homogeneous Markov chains, the generally admitted solution is to treat the model as a set of multinomial proportions and to compute confidence intervals accordingly. In the case of HMMs, there exists a theoretical solution using the Hessian matrix, but it cannot be applied on large samples, so we have to rely on simulated methods which can be very computationally intensive (Visser, Raijmakers & Molenaar, 2000). Finally, as far as we know, there is still no published method for the construction of confidence intervals in the case of the MTD model.

When comparing different models, global tests are not always sufficient. For instance, a likelihood ratio test or a measure such as the Bayesian Information Criterion (BIC, e.g. Raftery (1999)) can be used to discriminate between two models computed on the same data set. On the other hand, these methods cannot be used to compare models with the same structure, but computed on different samples, with possibly different sizes. This problem can be solved by the use of confidence intervals. We consider here two particular examples:

(a) We observed sequences of successive rainy (R) and sunny (S) days at two meteorological stations and we built a first order homogeneous Markov chain for each location. At station A, we observed the weather for 201 successive days and we obtained the following crosstable (C_A) and transition matrix (Q_A):

$$C_A = \begin{array}{c|cc|} & t & \\ t-1 & R & S \\ \hline R & 40 & 25 \\ S & 35 & 100 \\ \hline \end{array} \quad , \quad Q_A = \begin{array}{c|cc|} & t & \\ t-1 & R & S \\ \hline R & 0.62 & 0.38 \\ S & 0.26 & 0.74 \\ \hline \end{array}$$

At station B, we observed the weather for 301 successive days and we obtained the following crosstable and transition matrix:

$$C_B = \begin{array}{c|cc|} & t & \\ t-1 & R & S \\ \hline R & 45 & 40 \\ S & 50 & 165 \\ \hline \end{array} \quad , \quad Q_B = \begin{array}{c|cc|} & t & \\ t-1 & R & S \\ \hline R & 0.53 & 0.47 \\ S & 0.23 & 0.77 \\ \hline \end{array}$$

The question is to determine whether the two transition matrices can be considered as similar. Confidence intervals provide a good answer. Using the method developed in this paper and a type I error of 0.05, each transition probability can be replaced by a confidence interval:

$$Q_A \Longrightarrow \begin{array}{c|cc|} & t & \\ t-1 & R & S \\ \hline R & [0.48;0.76] & [0.24;0.52] \\ S & [0.16;0.36] & [0.64;0.84] \\ \hline \end{array}$$

$$Q_B \Longrightarrow \begin{array}{c|cc|} & t & \\ t-1 & R & S \\ \hline R & [0.41;0.65] & [0.35;0.59] \\ S & [0.15;0.31] & [0.69;0.85] \\ \hline \end{array}$$

Since the confidences intervals overlap for all four probabilities, we cannot reject the hypothesis that the first-order transition processes are identical at the two chosen locations.

(b) A young monkey was observed during one hour and its behavior was recorded each five seconds for a total of 720 data points. There are three possible behaviors: Passivity (Pa), Exploration (E), and Play (P). Two models were computed, an homogeneous first-order Markov chain, and a two hidden states HMM. According to BIC, the homogeneous Markov chain is clearly rejected in favor of the hidden model. The transition matrix between the two hidden states S_1 and S_2 is

$$A = \begin{array}{c|cc|} & t & \\ t-1 & S_1 & S_2 \\ \hline S_1 & 0.95 & 0.05 \\ S_2 & 0.10 & 0.90 \\ \hline \end{array}$$

and the probability distributions corresponding to the hidden states are

$$C_1 = (\ 0.21 \quad 0.14 \quad 0.65\ ) \quad \text{and} \quad C_2 = (\ 0.16 \quad 0.29 \quad 0.55\ )$$

We would like to know which of the three probabilities significantly differ between distributions C_1 and C_2. The computation of confidence intervals again solves the problem. Using the Viterbi algorithm (Forney, 1973), we determined that the subject spent two third of the time (480 data points) in the first hidden state and one third in the second one. This information was used to compute the following confidence intervals:

$$C_1 \Longrightarrow (\ [0.16; 0.26] \quad [0.09; 0.19] \quad [0.60; 0.70]\ )$$
$$C_2 \Longrightarrow (\ [0.09; 0.23] \quad [0.22; 0.36] \quad [0.48; 0.62]\ )$$

The first and third interval overlap, indicating that the corresponding probabilities of being in the Passive and Play behaviors are not statistically different in the two hidden states. The only significative difference concerns the proportion of time spent in the Explore behavior which is significantly higher in the second hidden state.

The paper is organized as follows: In Section 2 we recall the principle used to estimate sample size and to build confidence intervals in the case of a single multinomial distribution. In Sections 3 to 5, we show how this principle can be applied to different Markovian models. Numerical applications are provided in Section 6, and a Conclusion section ends the paper.

2. Simultaneous confidence intervals for a multinomial distribution

Let $(\theta_1, \theta_2, \ldots, \theta_c)$ denote the theoretical probability distribution of a categorical random variable Y taking values in $\{1, \ldots, c\}$. Let $(n_1, n_2, \ldots, n_c)$ be the observed frequency distribution computed from a sample of size $n = \sum_{i=1}^{c} n_i$, and let $(p_1, p_2, \ldots, p_c)$ denote the corresponding empirical probability distribution. The basic question is to know what the sample size should be in order to obtain a given precision of the estimation, or alternatively, what is the precision of the estimation given the current sample size. Mathematically, we have to find the minimal n such that

$$P\left(\bigcap_{i=1}^{c}(|p_i - \theta_i| \leq d)\right) \geq 1 - \alpha$$

where α is the type I error, and where d denotes the precision of the estimation, that is the maximal difference allowed between any element of the theoretical probability distribution and its empirical estimation, or equivalently the half-length confidence interval around the empirical estimations.

This problem has been adressed many times. See e.g. Fitzpatrick & Scott (1987), Thompson (1987), Sison & Glaz (1995) for some recent developments. Following Adcock (1997), a conservative rule is to have a sample size of at least

$$n = 0.25\frac{\chi^2_{[1,\alpha/c]}}{d^2}$$

where $\chi^2_{[1,\alpha/c]}$ is the threshold of a chi-square distribution with one degree of freedom and probability equal to α/c. This rule is very easy to use, but it suffers some shortcomings, among them the fact that the sample size increases monotonically with the number c of

categories. Thompson (1987) proposed to use instead

$$n = \max_{c \geq 1} \left\{ \frac{\frac{1}{c}\left(1 - \frac{1}{c}\right) \chi^2_{[1,\alpha/c]}}{d^2} \right\} \tag{1}$$

Even if this rule is not perfect, its degree of conservatism being unknown, it presents the advantage to be independent from the number of categories, the maximum of n occuring at low values of c.

Thompson provided a table giving the quantity $d^2 n$ for different values of α: $d^2 n$=1.00635 for $\alpha = 10\%$, 1.27359 for $\alpha = 5\%$, and 1.96986 for $\alpha = 1\%$. Dividing this value by the desired d^2 gives the required sample size, whatever the number of categories of the distribution. For instance, $d = 0.1$ and $\alpha = 5\%$ lead to $n = 1.27359/0.1^2 = 127.359$, so a sample of at least 128 data points should be used.

It is also possible to consider the reverse problem and to compute the actual half confidence interval d given the number of data used for the estimation. If we have for instance $n = 300$ data points and we use $\alpha = 5\%$, the precision is then equal to $d = \sqrt{1.27359/300} = 0.0652$, what means that all probabilities of the multinomial distribution have been estimated with a precision of ± 0.0652. Notice that when the number of data points n is very small, Thompson's formula can provide results larger than one.

In the next sections, we will use Thompson's principle as the reference method for confidence intervals computing. The reason for that choice is the good tradeoff between the performance of the method and its easiness of implementation. However, it could be replaced by another principle without any alteration of the core material of this paper.

3. Homogeneous Markov chains

We turn now to the more general problem of an homogeneous Markov chain of order $f \geq 1$. We do not consider here the special case of an order zero Markov chain, that is a memoryless model or independence model, because this model reduces to the single multinomial distribution case considered in the previous section. Let Y_t be a categorical random variable taking values in $\{1, \ldots, c\}$ and observed for $n + f$ successive periods. In an order f homogeneous Markov chain, the present is explained by the combination of the last f observations. This model is characterized by a transition matrix of size $c^f \times c^f$. Given the set of constraints applying on chains of order higher than one, only c probabilities can be non-zero on each row, and they constitute a multinomial probability distribution. It is then possible to apply the principle defined in the previous section to each of these distributions, the only requirement being the knowledge of the number of data used to compute each row from the data set. See e.g. Kemeny & Snell (1976) or Brémaud (1999) for a complete treatment of Markov chain theory.

Since each row of a Markov chain contains a maximum of c non-zero elements, transition matrices associated with Markov chains of various orders really differ only by their number of rows. We consider here the first order model, the generalization to higher orders being straightforward. Let

$$Q = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & \cdots & p_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ p_{c1} & p_{c2} & \cdots & p_{cc} \end{pmatrix}$$

be the transition matrix associated to a first-order Markov chain, with $\sum_{j=1}^{c} p_{ij} = 1$, $i = 1, \ldots, c$, and let

$$C = \begin{pmatrix} n_{11} & n_{12} & \ldots & n_{1c} \\ n_{21} & n_{22} & \ldots & n_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ n_{c1} & n_{c2} & \ldots & n_{cc} \end{pmatrix}$$

be the corresponding crosstable, $\sum_{i=1}^{c} \sum_{j=1}^{c} n_{ij} = n$. Let $n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$ denote the number of data points on the i-th row in C. The quantity $n_{i\cdot}$ is also the number of data points used to compute the i-th row in Q. Applying Thompson's principle on $n_{1\cdot}, \ldots, n_{c\cdot}$, we obtain confidence intervals for each parameter of the model.

In a conservative approach, we could consider the entire transition matrix Q to be reliably estimated if and only if each row is reliably estimated. We would just have then to consider the minimum of the row totals $n_{1\cdot}, \ldots, n_{c\cdot}$ and to check whether this value is at least equal to the minimal sample size given by equation (1), or to compute the empirical precision d associated with this value.

4. Mixture Transition Distribution models

In many cases, the use of high-order Markov chains is not possible, because the number of parameters increases exponentially with the dependence order. Different modelings have been proposed to reduce this problem. Among them, the Mixture Transition Distribution (MTD) model introduced by Raftery (1985) proved to be very interesting, combining parsimony with quality of modeling (Berchtold & Raftery, 2002). Let Y_t be a categorical random variable taking values in $\{1, \ldots, c\}$. The principle of the MTD model is to consider independently the influence of each lag upon the present. The basic equation is

$$\begin{aligned} P(Y_t = i_0 | Y_{t-f} = i_f, \ldots, Y_{t-1} = i_1) &= \sum_{g=1}^{f} \lambda_g \, P(Y_t = i_0 | Y_{t-g} = i_g) \\ &= \sum_{g=1}^{f} \lambda_g \, q_{i_g i_0} \end{aligned}$$

where f is the order of the Markov chain, $\Lambda = (\lambda_1, \ldots, \lambda_f)$, $\sum_i \lambda_i = 1$, is a vector of lag weights, and where

$$Q = \begin{pmatrix} q_{11} & \ldots & q_{1c} \\ \vdots & & \vdots \\ q_{c1} & \ldots & q_{cc} \end{pmatrix}$$

is a transition matrix of the same size as the transition matrix associated with the first-order homogeneous Markov chain. The model is very parsimonious since each new lag adds only one parameter to the model.

The MTD model is estimated using either a dedicated algorithm such as the one proposed in Berchtold (2001), or more general methods such as the Expectation-Maximization (EM) algorithm (see e.g. McLachlan & Krishnan (1996)) or a genetic algorithm (Holland, 1975).

The parameters of the MTD model can be decomposed into $c+1$ different distributions: the c rows of the transition matrix Q and the vector of weights Λ. When the elements of Λ are constrained to be non-negative, the most current case, the $c+1$ sets of parameters are all probability distributions. The only requirement for applying equation (1) is then to estimate the number of data used to compute each of these $c+1$ distributions. That can be done using a principle similar to the E step of an EM algorithm. Suppose that we have $n+f$ successive observations of the random variable numbered from Y_{-t+1} to Y_n. Suppose that at each time t, the probability distribution was influenced by one of the f lags and let Z_t be a size f vector indicating which lag was used at that time:

$$Z_t(g) = \begin{cases} 1\,, & \text{if lag } g \text{ explains } X_t \\ 0\,, & \text{otherwise} \end{cases}$$

The lag really used to explain Y_t being unobserved, we say that the model includes a hidden component. The expectation of $Z_t(g)$ is computed as

$$\begin{aligned} \tilde{z}_t(g) &= E(Z_t(g)|Y_t, \Lambda, Q) \\ &= P(z_t(g)=1|y_t = i_0, \Lambda, Q) \\ &= \frac{P(y_t = i_0|z_t(g)=1)\, P(z_t(g)=1)}{\sum_{k=1}^{f} P(y_t = i_0|z_t(k)=1)\, P(z_t(k)=1)} \\ &= \frac{\lambda_{g,t}\, q_{i_g i_0}}{\sum_{k=1}^{f} \lambda_k\, q_{i_k i_0}} \end{aligned}$$

The weight coefficient for the g-th lag is then estimated as

$$\hat{\lambda}_g = \frac{\sum_{t=1}^{n} \tilde{z}_t(g)}{n}$$

and the total number of data points used to compute the Λ vector is

$$\sum_{g=1}^{f}\sum_{t=1}^{n} \tilde{z}_t(g) = n$$

This quantity is then used to compute confidence intervals for the weight parameters λ_g.

The computation of confidence intervals for the parameters of the transition matrix Q is more difficult. In a MTD model, each probability q_{ki_0} in Q corresponds to f different situations occuring in the data: $\{(y_{t-f}=k, y_t=i_0),\ (y_{t-f+1}=k, y_t=i_0), \ldots, (y_{t-1}=k, y_t=i_0)\}$. Let $\hat{N}$ be a $(c \times c)$ matrix whose element (k, i_0) is defined as

$$\hat{n}_{ki_0} = \sum_{\substack{t=1 \\ y_t=i_0}}^{n} \quad \sum_{\substack{g=1 \\ y_{t-g}=k}}^{f} \tilde{z}_t(g)$$

This matrix contains an estimation of the number of data used to compute each element of Q. Each row total in $\hat{N}$ is then used to build confidence intervals for the probabilities of the corresponding row in Q.

A MTD model can be used to build an estimation of the high-order transition matrix of the corresponding homogeneous Markov chain. The number of data points used to compute the probability corresponding to $Y_{t-f} = i_f, \ldots, Y_t = i_0$ is estimated by

$$\hat{m}_{i_f,\ldots,i_0} = \sum_{g=1}^{f} \lambda_g \hat{n}_{i_g i_0}$$

and the total number of data points used for the row defined by $Y_{t-f} = i_f, \ldots, Y_{t-1} = i_1$ is

$$\hat{m}_{i_f,\ldots,\cdot} = \sum_{j=1}^{c} \hat{m}_{i_f,\ldots,i_0}$$

The $\hat{m}_{i_f,\ldots,\cdot}$ are then used to compute confidence intervals for each high-order transition probability using Thompson's principle.

The basic MTD model allows many extensions, among them the MTDg model and the use of covariates. In a MTDg model, a different matrix is used to represent the transition process between each lag of the model and the present. There are then f transition matrices $\hat{Q}_1, \ldots, \hat{Q}_f$ and f corresponding matrices $\hat{N}^{(1)}, \ldots, \hat{N}^{(f)}$, the element (k, i_0) of $\hat{N}^{(g)}$ being defined as

$$\hat{n}^{(g)}_{ki_0} = \sum_{\substack{t=1 \\ y_t = i_0 \\ y_{t-g} = k}}^{n} \tilde{z}_t(g)$$

The row totals of $\hat{N}^{(g)}$ are used to build confidence intervals around the parameters of the transition matrix Q_g. The number of data points used to compute the probability defined by $Y_{t-f} = i_f, \ldots, Y_t = i_0$ in the f-th order transition matrix is estimated by

$$\hat{m}_{i_f,\ldots,i_0} = \sum_{g=1}^{f} \lambda_g \hat{n}^{(g)}_{i_g i_0}$$

A mixture model can also be used to add one or several categorical covariates to a Markovian model. Even if the resulting model is no more a Markov chain, the above principles and formulas stay valid, the difference being that in addition to the matrix Q, we have to consider a set of transition matrices giving the relation between each covariate and the random variable Y_t. For instance, if we add a covariate V taking b different values to an order f MTD model, the vector of weights becomes

$$\Lambda = (\lambda_1, \ldots, \lambda_f, \lambda_v) \ , \quad \lambda_1 + \ldots + \lambda_f + \lambda_v = 1$$

where λ_v is the weight of the covariate in the mixture. Vectors $Z_{g,t}$ are also of size $f + 1$, the last component representing the fact that the covariate was actually responsible for the observed value of Y_t. A matrix of size $(b \times c)$, whose (k, i_0)-th element is

$$\hat{n}^{(f+1)}_{ki_0} = \sum_{\substack{t=1 \\ y_t = i_0 \\ v_t = k}}^{n} \tilde{z}_t(f+1)$$

is estimated for the covariate, and the number of data points used to compute the element of the high-order matrix defined by $V = k, Y_{t-f} = i_f, \ldots, Y_t = i_0$ is then equal to

$$\hat{m}_{k,i_f,\ldots,i_0} = \sum_{g=1}^{f} \lambda_g \hat{n}^{(g)}_{i_g i_0} + \lambda_v \hat{n}^{(f+1)}_{k i_0}$$

5. Hidden Markov Models

In a Hidden Markov Model (HMM), the state taken by the Markov chain at time t is unknown. We observe instead a second random variable whose distribution depends on the state taken by the hidden chain. Let X_t be a categorical random variable taking values in $\{1, \ldots, K\}$ and governed by a hidden Markov chain of order ℓ. Let Y_t be a categorical random variable taking values in $\{1, \ldots, c\}$. To each of the K states of X_t correspond a different probability distribution for Y_t, the distribution actually used at time t being unknown. See e.g. Rabiner (1989) or MacDonald & Zucchini (1997) for a comprehensive treatment of HMMs.

We consider a first-order hidden transition matrix and n successive observations of Y_t numbered from Y_1 to Y_n. Three different sets of probabilities are used to fully identify a model: the unconditional distribution of the first hidden state, denoted by π, the transition matrix between hidden states, A, and the distributions of the random variable Y_t given each of the K hidden states, $C_1, \ldots, C_K$.

Theoretically, assuming the asymptotic normality of the parameter maximum likelihood estimators around their true value, it is possible to compute an exact confidence interval for each parameter of a HMM (Seber & Wild, 1989). In practice, however, it is not possible to use this method when the series of interest is more than about n=100 data points long (Visser, Raijmakers & Molenaar, 2000). Three alternative methods have been introduced in the last years: finite-differences approximation (Dennis & Schnabel, 1983), bootstrap (Efron & Tibshirani, 1986), and likelihood profiles (Meeker & Escobar, 1995). These methods present the advantage of being usable on long time series. On the other hand, as noted in Visser, Raijmakers & Molenaar (2000), results can be very sensible to truncations and rounding errors. Moreover, the required computations are very intensive. In this section, we show how the method used for homogeneous Markov chain and MTD models can be applied to hidden models at a very low cost.

The percentage of the time the model is using each of the K possible distributions for the observed random variable is unknown. However, this percentage can be estimated as follows: HMM parameters are estimated using the Baum-Welch algorithm (Rabiner, 1989), a customized version of the Expectation-Maximization (EM) algorithm. Similarly to what we did for the MTD model, we suppose that at each time t, one of the K possible distributions was used. Let Z_t be a size K vector indicating which distribution was used at time t:

$$Z_t(g) = \begin{cases} 1\,, & \text{if distribution } C_g \text{ was active at time } t \\ 0\,, & \text{otherwise} \end{cases}$$

During the E step of the estimation algorithm, the expectation of $Z_t(g)$ is computed as follows. First, we define two auxiliary quantities:

$$\alpha_t(g) = P(Y_1, \ldots, Y_t, X_t = g | \pi, A, C_1, \ldots, C_K)\,, \quad t = 1, \ldots, n$$

$$\beta_t(g) = P(Y_{t+1}, \ldots, Y_n | X_t = g, \pi, A, C_1, \ldots, C_K) \ , \quad t = 1, \ldots, n$$

Using $\alpha_1(g) = \pi(g) C_g(y_1)$ and $\beta_n(g) = 1$, $g = 1, \ldots, K$, to initialize the computation, we obtain iteratively

$$\alpha_{t+1}(g) = \left(\sum_{i=1}^{K} \alpha_t(i) a(i,g) \right) C_g(y_{t+1}) \ , \quad t = 2, \ldots, n$$

$$\beta_t(g) = \sum_{i=1}^{K} C_g(y_{t+1}) \beta_{t+1}(g) a(i,g) \ , \quad t = n-1, \ldots, 1$$

The expectation of $Z_t(g)$ is then

$$\begin{aligned} \tilde{z}_t(g) &= E(Z_t(g) | Y_t, \pi, A, C_1, \ldots, C_K) \\ &= P(z_t(g) = 1 | y_t, \pi, A, C_1, \ldots, C_K) \\ &= \frac{\alpha_t(g) \, \beta_t(g)}{L(Y_1, \ldots, Y_n)} \end{aligned}$$

where $L(Y_1, \ldots, Y_n) = \sum_{i=1}^{K} \alpha_n(i)$ is the likelihood of the observed data. The total number of data points used to compute C_g is estimated as

$$\sum_{t=1}^{n} \tilde{z}_t(g)$$

and this quantity is used to compute confidence intervals for the probabilities in C_g.

By definition of the model, each observation of Y_t is directly linked to a hidden state, so the length of the unobserved sequence of hidden states is n. The probability distribution of the first hidden state is given by $(\tilde{z}_1(1), \ldots, \tilde{z}_1(K))$, and the sum of this vector, one, is the number of data used to compute π. Obviously, one data point is not enough to obtain a reliable estimation, but in practice, when only one sample is used to identify the model, the precise knowledge of the active state at the beginning of the series is generally of no use. On the other hand, when a same HMM is computed from several independent samples, the total number of data points used to compute each vector Z_t, and in particular Z_1, is equal to the number of independent samples, what can lead to a reliable estimation of π.

The number of data entering in the computation of the g-th row of the hidden transition matrix A is estimated as

$$\sum_{t=1}^{n-1} \tilde{z}_t(g)$$

and this quantity is used to estimate the precision of the g-th row of A. Notice that the summation is taken up to $n-1$ rather than n, because the transition matrix expresses the relation between two successive hidden states, the first in row and the second in column. A summation up to n is then not possible, because there is no data point after Y_n.

We turn now to the case of a high-order HMM, that is a HMM with a transition matrix A of order $\ell > 1$. In addition to the computation of the first hidden state distribution, π, the complete identification of the model requires the computation of the distribution of the second hidden state given the first one, $\pi_{2|1}$, the distribution of the third hidden state given the first two, $\pi_{3|1,2}$, and so on until $\pi_{\ell|1,\ldots,\ell-1}$. In the Baum-Welch algorithm, the vectors $\tilde{Z}_t$ are replaced by the computation of matrices giving the expectation of observing ℓ successive hidden states and we define:

$$\tilde{z}_t(i_{\ell-1},\ldots,i_0) = P(X_{t-\ell+1} = i_{\ell-1},\ldots,X_t = i_0|Y_1,\ldots,Y_t)\ , \quad t = \ell,\ldots,n \tag{2}$$

For $t < \ell$, we define matrices giving the expectation of observing t successive hidden states:

$$\tilde{z}_t(i_{t-1},\ldots,i_0) = P(X_1 = i_{t-1},\ldots,X_t = i_0|Y_1,\ldots,Y_t)\ , \quad t = 1,\ldots,\ell-1 \tag{3}$$

The number of data points used to compute a probability distribution in the model is estimated by summing up corresponding elements in $\tilde{z}_t(i_{\ell-1},\ldots,i_0)$ and in $\tilde{z}_t(i_{t-1},\ldots,i_0)$. For instance, the number of data points corresponding to the first row of A, which is defined by $i_{\ell-1} = \ldots = i_0 = 1$, is computed as

$$\sum_{\substack{t=\ell \\ i_{\ell-1}=\ldots=i_0=1}}^{n-1} \tilde{z}_{i_{\ell-1},\ldots,i_0}(t)$$

and the number of data points used to compute the distribution C_1 is estimated as

$$\sum_{\substack{t=1 \\ i_0=1}}^{\ell-1} \tilde{z}_{i_{t-1},\ldots,i_0}(t) + \sum_{\substack{t=\ell \\ i_0=1}}^{n} \tilde{z}_{i_{\ell-1},\ldots,i_0}(t)$$

The Double Chain Markov Model (DCMM) first introduced by Paliwal (1993) in the context of speech recognition and then developed and generalized by Berchtold (1999, 2002) is an extension of the basic HMM. In a DCMM, the hypothesis of conditional independence between successive observations of the random variable Y_t is removed and the relation between observations is instead modeled by the mean of a Markov chain. The probability distributions $C_1,\ldots,C_K$ are then replaced by K transition matrices, $Q_1,\ldots,Q_K$, possibly of order $f \geq 1$. This model proved to be very useful especially for the modeling of animal behavior (Berchtold & Sackett, 2002).

For technical reasons (Berchtold, 2002), the estimation of a DCMM requires f initial observations of the visible process numbered from Y_{-f+1} to Y_0. These observations are used only to initialize the estimation process and they do not enter in the computation of the likelihood. By consequence, the expectations of equations (2) and (3) are conditioned on $Y_{-f+1},\ldots,Y_t$ instead of $Y_1,\ldots,Y_t$. The number of data points used to compute the element of the matrix Q_g defined by $Y_{t-f+1} = j_{f-1},\ldots,Y_t = j_0, Y_{t+1} = j$ is given by

$$\sum_{\substack{t=1 \\ i_0=g \\ Y_{t-f+1}=j_{f-1},\ldots,Y_t=j_0,Y_{t+1}=j}}^{\ell-1} \tilde{z}_{i_{t-1},\ldots,i_0}(t) + \sum_{\substack{t=\ell \\ i_0=g \\ Y_{t-f+1}=j_{f-1},\ldots,Y_t=j_0,Y_{t+1}=j}}^{n-1} \tilde{z}_{i_{\ell-1},\ldots,i_0}(t)$$

and the total number of data points used to compute the row of the matrix Q_g defined by $Y_{t-f+1} = j_{f-1}, \ldots, Y_t = j_0$ is then given by

$$\sum_{\substack{t=1 \\ i_0=g \\ Y_{t-f+1}=j_{f-1},\ldots,Y_t=j_0}}^{\ell-1} \tilde{z}_{i_{t-1},\ldots,i_0}(t) + \sum_{\substack{t=\ell \\ i_0=g \\ Y_{t-f+1}=j_{f-1},\ldots,Y_t=j_0}}^{n-1} \tilde{z}_{i_{\ell-1},\ldots,i_0}(t)$$

As proposed in Berchtold (2002), it is possible to replace the high-order hidden and/or visible transition matrices of a HMM or of a DCMM by a MTD model. In that case, confidence intervals are computed by combining equations provided in Sections 4 and 5.

6. Numerical applications

Two numerical applications are proposed in this section. In the first one, we consider different Markovian approaches for the modeling of a sequence of wind speed measurements, and confidence intervals are computed in each case. In the second one, we compare confidence intervals for a reference HMM with previously published results obtained from other methods.

6.1. Wind speed data set

We use here a sample of 672 successive hourly measures of the wind speed at Belmullet, Ireland (source: Raftery (1985)). We consider four different speeds ranging from 1: no wind, to 4: excessively high wind. All models are computed on data points 5 to 672, conditionally on the first four observations. In each case, the type I error is set to α=0.05.

Overall, using the Bayesian Information Criterion, the preferred model is the second-order MTD. On the other hand, the independence model and the two-hidden state HMM (model of conditional independence) are clearly rejected, showing a strong dependence between successive wind speed observations. However, we provide hereafter the parameters of all models with their repective half-length confidence intervals.

6.1.1. Independence model

The empirical probability distribution of the four wind speeds is

$$(0.1931 \quad 0.4296 \quad 0.2485 \quad 0.1287)$$

Using Thompson's principle, the corresponding precision is then $d = \sqrt{1.27359/668} = 0.0437$, what leads to the following confidence intervals:

$$(\ [0.1494; 0.2368] \quad [0.3859; 0.4733] \quad [0.2048; 0.2922] \quad [0.0850; 0.1724]\ )$$

Here, wind speed 2 has a probability larger than any other wind speed, and wind speed 4 has a lower probability than wind speed 3.

Using the reverse computation, the obtention of a precision $d \leq 0.01$ would have required at least $n = 1.27359/(0.01^2) = 12'740$ data points. On the other hand, $n = 128$ data points would have been sufficient for a precision of 0.1.

6.1.2. Homogeneous Markov chains

The crosstable associated to the transition matrix of the first-order homogeneous Markov chain is written

	t				
$t-1$	1	2	3	4	$n_{i\cdot}$
1	103	27	0	0	130
2	26	226	34	0	286
3	0	34	119	13	166
4	0	0	13	73	86

and the corresponding transition matrix is written

	t				
$t-1$	1	2	3	4	d_i
1	0.79	0.21	0	0	0.0990
2	0.09	0.79	0.12	0	0.0667
3	0	0.20	0.72	0.08	0.0876
4	0	0	0.15	0.85	0.1217

where d_i is the precision or half-length confidence interval associated to each probability of the i-th row. The four precisions vary much, because of the large discrepancy in the number of data points available for the computation of each row (from 86 for row 4 to 286 for row 2). In a conservative approach, the overall precision of the estimation is ± 0.1217, since this is the precision of the worst estimated row. We can also compute the number of data points which would be necessary to obtain a given precision. If we choose a precision $d = 0.07$, then we should have at least $n_i = \sqrt{1.27359/(0.07)^2} \approx 260$ data points per row to ensure this precision, what is not the case here, only the second row being computed with such a number of data points.

In the case of the second order homogeneous Markov chain, we obtain the following results, the transition matrix being written in reduced form:

		t				
$t-2$	$t-1$	1	2	3	4	d_i
1	1	0.81	0.19	0	0	0.1107
2	1	0.73	0.27	0	0	0.2213
3	1	0	0	0	0	–
4	1	0	0	0	0	–
1	2	0.33	0.63	0.04	0	0.2172
2	2	0.07	0.81	0.12	0	0.0752
3	2	0.03	0.76	0.21	0	0.1935
4	2	0	0	0	0	–
1	3	0	0	0	0	–
2	3	0	0.53	0.47	0	0.1935
3	3	0	0.13	0.81	0.06	0.1035
4	3	0	0	0.54	0.46	0.3130
1	4	0	0	0	0	–
2	4	0	0	0	0	–
3	4	0	0	0.31	0.69	0.3130
4	4	0	0	0.12	0.88	0.1321

Some transitions were not observed in the dataset, so several rows and their associated precision could not be computed. Precisions range here from ± 0.0752 to ± 0.3130. Obviously, since there are much more parameters to estimate here, these half-length confidence intervals are larger in average than the half-length intervals computed for the first-order transition matrix.

6.1.3. MTD model

We used a second-order MTD model to approximate the real second-order homogeneous Markov chain. The weights associated to each lag are $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$ for a precision of ± 0.0437. The transition matrix of the MTD model is written

$$Q = \begin{array}{c|cccc|c} & t & & & & \\ t-1 & 1 & 2 & 3 & 4 & d_i \\ \hline 1 & 0.82 & 0.18 & 0 & 0 & 0.0976 \\ 2 & 0.07 & 0.83 & 0.10 & 0 & 0.0664 \\ 3 & 0 & 0.15 & 0.80 & 0.05 & 0.0900 \\ 4 & 0 & 0 & 0.12 & 0.88 & 0.1200 \end{array}$$

and the second-order approximate transition matrix is written

		t				
$t-2$	$t-1$	1	2	3	4	d_i
1	1	0.82	0.18	0	0	0.0976
2	1	0.60	0.37	0.03	0	0.0841
3	1	0.58	0.17	0.24	0.01	0.0951
4	1	0.58	0.12	0.04	0.26	0.1029
1	2	0.29	0.64	0.07	0	0.0725
2	2	0.07	0.83	0.10	0	0.0664
3	2	0.05	0.63	0.31	0.01	0.0715
4	2	0.05	0.58	0.11	0.26	0.0746
1	3	0.25	0.16	0.56	0.03	0.0921
2	3	0.02	0.35	0.59	0.04	0.0805
3	3	0	0.15	0.80	0.05	0.0900
4	3	0	0.10	0.60	0.30	0.0965
1	4	0.25	0.05	0.08	0.62	0.1117
2	4	0.02	0.25	0.11	0.62	0.0927
3	4	0	0.05	0.32	0.63	0.1081
4	4	0	0	0.12	0.88	0.1200

The comparison between this approximate matrix and the real second-order transition matrix given previously reveals several very interesting points:

(a) Overall, the precision of the estimation is much better in the MTD model than for the real Markov chain. This is a consequence of the fact that the MTD model has less parameters and that each transition is used to represent different situations, leading to a much more efficient use of the available data points.

(b) Transition probabilities reliably estimated in the real second-order Markov chain do not change much in the MTD approximation. On the other hand, unreliable probabilities can be very different. This is an empirical justification of the usefulness of the MTD model.

(c) Since the MTD model suppresses the direct link between lags, probabilities, hence confidence intervals, can be computed even for transitions which do not occur in the sample (for instance on rows 3, 4, 8, 9, 13, and 14). This can be seen as a drawback of the model, but it is possible to simply ignore these rows.

6.1.4. HMM and DCMM

We consider first a HMM with two hidden states and a first-order dependence between hidden states. Using a standard EM algorithm, we obtained the following model: The distribution of the first hidden state is (0 1), what means that the process begins in the second hidden state. Since we used only one sequence of observations, this initial distribution was computed using one data point, and the corresponding precision is ±1.1285, what is of no use. The hidden transition matrix is given by

$$A = \begin{array}{c|cc|c} & t & & \\ t-1 & 1 & 2 & d_i \\ \hline 1 & 0.98 & 0.02 & 0.0535 \\ 2 & 0.04 & 0.96 & 0.0757 \end{array}$$

and the probability distributions of the wind speeds corresponding to the two hidden states are

$$C_1 = \begin{array}{c|cccc|c} 1 & 2 & 3 & 4 & d \\ \hline 0.29 & 0.64 & 0.07 & 0 & 0.0535 \end{array}$$

$$C_2 = \begin{array}{c|cccc|c} 1 & 2 & 3 & 4 & d \\ \hline 0 & 0.01 & 0.60 & 0.39 & 0.0757 \end{array}$$

The analysis of these results shows that the probability of observing wind speeds 1 and 2 is significantly higher in hidden state 1 than in hidden state 2, and that the probability of observing wind speeds 3 and 4 is significantly higher in state 2. Formulas for the computation of the number of data points on each row of A and for the number of data points in C_1 and C_2 being almost identical, the precision has obviously to be similar in the two cases.

We consider finally a DCMM with the same settings used before for the HMM, but with a first-order dependence between successive wind speed observations. As before, the estimation of the first hidden state distribution is (0 1) for a precision of ±1.1285. The hidden transition matrix is given by

$$A = \begin{array}{c|cc|c} & t & & \\ t-1 & 1 & 2 & d_i \\ \hline 1 & 0.90 & 0.10 & 0.0609 \\ 2 & 0.10 & 0.90 & 0.0628 \end{array}$$

and the two transition matrices corresponding to the visible process are

$$Q_1 = \begin{array}{c|cccc|c} & & t & & & \\ t-1 & 1 & 2 & 3 & 4 & d_i \\ \hline 1 & 0.69 & 0.31 & 0 & 0 & 0.1269 \\ 2 & 0.27 & 0.69 & 0.04 & 0 & 0.1159 \\ 3 & 0 & 0 & 0.89 & 0.11 & 0.1050 \\ 4 & 0 & 0 & 0.24 & 0.76 & 0.1531 \end{array}$$

and

$$Q_2 = \begin{array}{c|cccc|c} t-1 \backslash t & 1 & 2 & 3 & 4 & d_i \\ \hline 1 & 0.95 & 0.05 & 0 & 0 & 0.1581 \\ 2 & 0 & 0.84 & 0.16 & 0 & 0.0816 \\ 3 & 0 & 0.67 & 0.33 & 0 & 0.1588 \\ 4 & 0 & 0 & 0 & 1 & 0.2006 \end{array}$$

We see that the probability for wind speed three to reccur at time $t+1$ when active at time t is significantly higher in the first hidden state (0.89 ± 0.1050 against 0.33 ± 0.1588 for state 2). On the other hand, the probability of recurring for the three other wind speeds is never significantly different between the two hidden states.

In matrix Q_1, wind speeds 1 and 2 are transition states, when wind speeds 3 and 4 constitute a final class. So, it is impossible to go from high (3 and 4) to low (1 and 2) wind speeds using continuously this matrix. This type of transition imply to switch at some point to the transition matrix Q_2 associated with the second hidden state. Notice also the particular structure of matrix Q_2. This matrix is composed of two final classes: $\{1, 2, 3\}$ and $\{4\}$. Hidden state 2 is then active in two distinct situations: at low and medium wind speeds in one hand, and for excessively high wind on the other hand, without direct transition between these two situations.

6.2. HMM comparison study

Visser, Raijmakers & Molenaar (2000) used the following two-hidden states, three output values HMM model to generate a dataset of length 2000:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$$

$$C_1 = \begin{pmatrix} c_{11} & c_{12} & c_{13} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.0 & 0.4 \end{pmatrix}$$

$$C_2 = \begin{pmatrix} c_{21} & c_{22} & c_{23} \end{pmatrix} = \begin{pmatrix} 0.0 & 0.3 & 0.7 \end{pmatrix}$$

They estimated then the same HMM from the generated data set and they computed confidence intervals using likelihood profiles, 500 and 1000 samples bootstrap, and finite differences approximation of the Hessian. As discussed by Visser et al., confidence intervals obtained from likelihood profiles and bootstrap are very similar, when confidence intervals obtained from finite-differences approximation of the Hessian seem generally too narrow. For the purpose of comparison, we report in Table 1 the half-length confidence intervals obtained from 1000 samples bootstrap and from finite-differences approximation of the Hessian, along with the corresponding intervals obtained using our multinomial method.

As can be seen in Table 1, confidence intervals obtained from the multinomial method are generally close to the ones obtained from other methods. These results confirm that our approach is well-founded, since it is possible to obtain reliable confidence intervals at only a fraction of the computational cost of previous approaches.

7. Conclusion

The computation of confidence intervals for the parameters of Markovian models is difficult because of the possible large number of parameters involved, and because parts of

Table 1. Comparison of confidence intervals obtained using different methods. The first column gives the parameters of the HMM defined in Visser, Raijmakers & Molenaar (2000). The second column gives the estimated parameters obtained by maximization of the log-likelihood. The third and fourth column report the 5% half-length confidence intervals obtained by Visser et al. using respectively 1000 samples bootstrap and finite-differences approximation of the Hessian. The fifth column gives the 5% half-length confidence intervals obtained using our multinomial method.

Parameter	Estimation Value	Bootstrap CI	Approximated Hessian CI	Multinomial CI
a_{11}	0.8843	0.0280	0.0269	0.0299
a_{12}	0.1157	0.0280	0.0269	0.0299
a_{21}	0.2864	0.0501	0.0690	0.0472
a_{22}	0.7136	0.0501	0.0690	0.0472
c_{11}	0.6163	0.0379	0.0328	0.0299
c_{12}	0.0	-	-	-
c_{13}	0.3837	0.0379	0.0328	0.0299
c_{21}	0.0	-	-	-
c_{22}	0.3102	0.0520	0.0387	0.0471
c_{23}	0.6898	0.0520	0.0387	0.0471

these models are sometimes not directly observed (hidden models). However, these issues are tractable using appropriate principles and algorithms. First of all, the parameters of Markovian models can be decomposed into a set of probability distributions. Second, when some elements of a model are not directly observable, it is always possible to estimate their expectation of appearance. Putting these two elements together leads to an efficient method for the computation of confidence intervals.

We dealt here mainly with the estimation of the number of data points involved in the computation of each parameter of Markovian models. For sample size and confidence intervals computation, we relied on the principle proposed by Thompson. The advantage is that another principle than Thompson's one could be easily implemented using the results of this paper. Moreover, we did not consider the case where a different confidence interval is computed for each element of a probability distribution, instead of having the same interval for all probabilities of the distribution. Again, these confidence intervals could be implemented using our results.

Different other methods have been proposed for the construction of confidence intervals in the case of hidden Markov models. As mentioned in the Introduction, exact computation is only feasible on very short data sequences. Alternative methods (likelihood profiles, bootstrap, and finite-differences approximation of the Hessian) answer to this issue, but they all have in common to be very computational intensive, so their application on very large data sets is not always possible. On the other hand, our multinomial approach is very easy to apply, since the required computations rely mainly on one run of the E-step of an EM algorithm, a standard tool. The door is so open to efficient software implementation such as the one proposed in the last release of MARCH, a software for the analysis and computation of Markovian models (http://www.AndreBerchtold.com/march.html).

Empirical results suggest that confidence interval obtained from the multinomial approach are very close to the ones obtained from much more costly methods. Moreover, the multinomial approach presents the advantage of unifying confidence intervals for both homogeneous Markov chains and hidden models in a same theoretical framework.

Acknowledgments

We would like to thank Gilbert Ritschard for useful discussions about confidence intervals and model comparison, and Ingmar Visser for providing the dataset of Section 6.2.

References

Adcock, C.J. (1997) Sample size determination: a review. *Statistician*, 46 (2), 261-283.

Berchtold, A. (1999) The Double Chain Markov Model. *Communications in Statistics: Theory and Methods*, 28, 2569-2589.

Berchtold A. (2001) Estimation in the Mixture Transition Distribution Model. *Journal of Time Series Analysis*, 22, 379-397.

Berchtold, A. (2002) High-Order Extensions of the Double Chain Markov Model. *Stochastic Models*, 18 (2), 193-227.

Berchtold, A., Raftery, A.E. (2002) The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical Science*, 17 (3), 328-356.

Berchtold, A., Sackett, G. (2002) Markovian Models for the Developmental Study of Social Behavior. *American Journal of Primatology*, 58 (3), 149-167.

Brémaud, P. (1999) *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer, New York.

Dennis, J.E., Schnabel, R.B. (1983) *Numerical methods for unconstrained optimization.* Prentice Hall, Englewood Clifff.

Efron, B., Tibshirani, R.J. (1986) Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science*, 1, 54-77.

Fitzpatrick, S., Scott, A. (1987) Quick Simultaneous Confidence Intervals for Multinomial Proportions. *Journal of the American Statistical Association*, 82, 875-878.

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

Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor.

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

MacDonald, I.L., W. Zucchini (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series.* Chapman & Hall, London.

McLachlan, G.J., Krishnan, T. (1996) *The EM Algorithm and Extensions.* John Wiley & Sons, New York.

Meeker, W.Q., Escobar, L.A. (1995) Teaching about approximate confidence regions based on maximum likelihood approximation. *American Statistician*, 49, 48-53.

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

Pegram, G.G.S. (1980) An Autoregressive Model for Multilag Markov Chains. *Journal of Applied Probability*, 17, 350-362.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2), 257-286.

Raftery, A.E. (1985) A model for high-order Markov chains. *Journal of the Royal Statistical Society B*, 47 (3), 528-539.

Raftery, A.E. (1999) Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods and Research*, 27, 411-427.

Seber, G.A.F., Wild, C.J. (1989) *Nonlinear regression*. John Wiley & Sons, New York.

Sison, C.P., Glaz, J. (1995) Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions. *Journal of the American Statistical Association*, 90, 366-369.

Thompson, S.K. (1987) Sample size for estimating multinomial proportions. *American Statistician*, 41, 42-46.

Visser, I., Raijmakers, M.E.J, Molenaar, P.C.M. (2000) Confidence intervals for hidden Markov model parameters. *British Journal of Mathematical and Statistical Psychology*, 53, 317-327.