Statistical Science
2002, Vol. 17, No. 3, 328–356

The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series

André Berchtold and Adrian E. Raftery

Abstract. The mixture transition distribution model (MTD) was introduced in 1985 by Raftery for the modeling of high-order Markov chains with a finite state space. Since then it has been generalized and successfully applied to a range of situations, including the analysis of wind directions, DNA sequences and social behavior. Here we review the MTD model and the developments since 1985. We first introduce the basic principle and then we present several extensions, including general state spaces and spatial statistics. Following that, we review methods for estimating the model parameters. Finally, a review of different types of applications shows the practical interest of the MTD model.

Key words and phrases: Mixture transition distribution (MTD) model, Markov chains, high-order dependences, time series, GMTD model, EM algorithm, spatial statistics, DNA, social behavior, wind, financial time series.

CONTENTS

André Berchtold is Assistant Professor in Statistics, Institute of Applied Mathematics, University of Lausanne, BFSH2, CH-1015 Lausanne, Switzerland (e-mail: Andre.Berchtold@imaa.unil.ch; web: www.andreberchtold.com). Adrian Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195-4322 (e-mail: raftery@stat.washington.edu; web: www.stat.washington.edu/raftery).

1. INTRODUCTION

This paper presents a review of the mixture transition distribution (MTD) model introduced by Raftery (1985a) for the modeling of high-order Markov chains. In this first section, we introduce the concept of Markov chains and the MTD model, and we motivate their use using two examples. This section can be used as a self-contained introduction to the subject for those who are interested in a brief overview. In the rest of the article we go into more detail.

In Section 2, we define the basic MTD model and we give its properties. Some extensions and generalizations are presented in Section 3, and parameter estimation is considered in Section 4. Section 5 presents a selection of applications for which the MTD model has proved to be useful. These applications come from different fields, including the analysis of wind directions, social behavior, DNA sequences and financial time series. Finally, we discuss some other approaches to the modeling of discrete-valued and non-Gaussian time series in Section 6.

1.1 Markov Chains

The Markov chain is a probabilistic model used to represent dependences between successive observations of a random variable. This model was introduced by Andrej Andreevic Markov at the beginning of the 20th century and it is used in many disciplines, including meteorology, geography, biology, chemistry, physics, behavior, social sciences and music. For comprehensive treatments of Markov chains and their applications, see, for example, Kemeny and Snell (1976), Kemeny, Snell and Knapp (1976), Karlin and Taylor (1981) and Brémaud (1999).

In this paper, we consider a discrete-time random variable X_t taking values in the finite set $\{1, \ldots, m\}$. Our goal is to predict or explain the value taken by X_t as a function of the values taken by previous observations of this same variable. The first-order Markov hypothesis says that the present observation at time t is conditionally independent of those up to and including time $(t-2)$ given the immediate past [time $(t-1)$]. Thus we can write

$$\begin{aligned} &P(X_t = i_0 | X_0 = i_t, \ldots, X_{t-1} = i_1) \\ &\quad = P(X_t = i_0 | X_{t-1} = i_1) \\ &\quad = q_{i_1 i_0}(t), \end{aligned}$$

where $i_t, \ldots, i_0 \in \{1, \ldots, m\}$. If we suppose that the probability $q_{i_1 i_0}(t)$ is time-invariant, it is replaced by $q_{i_1 i_0}$ and we have a *homogeneous* Markov chain. Considering all combinations of i_1 and i_0, we construct a transition matrix Q, each of whose rows sums to 1:

$$Q = \begin{array}{c} \\ X_{t-1} \\ 1 \\ \vdots \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} X_t & & & \\ 1 & \ldots & \ldots & m \end{array} \\ \begin{bmatrix} q_{11} & \ldots & \ldots & q_{1m} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ q_{m1} & \ldots & \ldots & q_{mm} \end{bmatrix} \end{array}.$$

Let

$$\chi_t = \big(x_t(1), \ldots, x_t(m)\big)' \tag{1}$$

be a vector such that $x_t(i) = 1$ if $X_t = i$ and 0 otherwise, and let $\hat{\chi}_t$ be the probability vector

$$\hat{\chi}_t = \big(P(X_t = 1), \ldots, P(X_t = m)\big)'. \tag{2}$$

Then the following relationships hold:

$$\hat{\chi}_t' = \chi_{t-1}' Q, \tag{3}$$

$$\hat{\chi}_t' = \chi_0' Q^t. \tag{4}$$

The process is fully defined once we know the initial vector χ_0 and the transition matrix Q.

In some situations, the present depends not only on the first lag, but on the last ℓ observations. We have then an ℓth-order Markov chain whose transition probabilities are

$$\begin{aligned} &P(X_t = i_0 | X_0 = i_t, \ldots, X_{t-1} = i_1) \\ &\quad = P(X_t = i_0 | X_{t-\ell} = i_\ell, \ldots, X_{t-1} = i_1) \\ &\quad = q_{i_\ell \cdots i_0}. \end{aligned} \tag{5}$$

For instance, if we set $\ell = 2$ and $m = 3$, the corresponding transition matrix is

$$Q = \begin{array}{cc} & \\ X_{t-2} & X_{t-1} \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 1 & 2 \\ 2 & 2 \\ 3 & 2 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{array} \begin{array}{c} \begin{array}{ccccccccc} X_t & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 3 \end{array} \\ \begin{array}{ccccccccc} X_{t-1} & 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} q_{111} & 0 & 0 & q_{112} & 0 & 0 & q_{113} & 0 & 0 \\ q_{211} & 0 & 0 & q_{212} & 0 & 0 & q_{213} & 0 & 0 \\ q_{311} & 0 & 0 & q_{312} & 0 & 0 & q_{313} & 0 & 0 \\ 0 & q_{121} & 0 & 0 & q_{122} & 0 & 0 & q_{123} & 0 \\ 0 & q_{221} & 0 & 0 & q_{222} & 0 & 0 & q_{223} & 0 \\ 0 & q_{321} & 0 & 0 & q_{322} & 0 & 0 & q_{323} & 0 \\ 0 & 0 & q_{131} & 0 & 0 & q_{132} & 0 & 0 & q_{133} \\ 0 & 0 & q_{231} & 0 & 0 & q_{232} & 0 & 0 & q_{233} \\ 0 & 0 & q_{331} & 0 & 0 & q_{332} & 0 & 0 & q_{333} \end{bmatrix} \end{array}.$$

When the order is greater than 1, notice that the transition matrix Q contains several elements corresponding to transitions that cannot occur. For instance, it is impossible to go from the row defined by $X_{t-2} = 1$ and $X_{t-1} = 2$ to the column defined by $X_{t-1} = 1$ and $X_t = 1$ because of the different

value taken by X_{t-1}. The probability of this transition is then 0 and we call this element a *structural zero*. Since the exact form of the transition matrix is known for any combination of ℓ and m, it is possible to rewrite Q in a more compact form excluding the structural zeros. This way of writing Q, as given by Pegram (1980), is called the *collapsed* or *reduced* form of Q and is denoted by R. The reduced form of the matrix corresponding to $\ell = 2$ and $m = 3$ is

$$R = \begin{array}{cc} & \\ X_{t-2} & X_{t-1} \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 1 & 2 \\ 2 & 2 \\ 3 & 2 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{array} \begin{array}{c} X_t \\ \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} q_{111} & q_{112} & q_{113} \\ q_{211} & q_{212} & q_{213} \\ q_{311} & q_{312} & q_{313} \\ q_{121} & q_{122} & q_{123} \\ q_{221} & q_{222} & q_{223} \\ q_{321} & q_{322} & q_{323} \\ q_{131} & q_{132} & q_{133} \\ q_{231} & q_{232} & q_{233} \\ q_{331} & q_{332} & q_{333} \end{bmatrix} \end{array}.$$

Each possible combination of ℓ successive observations of the random variable X is called a *state* of the model. The number of states is equal to m^ℓ ($= 3^2 = 9$ in our example). In the case of a first-order Markov chain, each value taken by the random variable X is also a state of the model.

The relationships (3) and (4) defined in the case of a first-order Markov chain still hold. Whatever the order is, there are $(m-1)$ independent probabilities in each row of the matrix Q, the last one of which is completely determined by the others since each row is a probability distribution summing to 1. The total number of independent parameters to be estimated is thus equal to $m^\ell(m-1)$. Given a set of observations, these parameters can be computed as follows. Let $n_{i_\ell \cdots i_0}$ denote the number of transitions of the type

$$X_{t-\ell} = i_\ell, \quad \ldots, \quad X_{t-1} = i_1, \quad X_t = i_0$$

in the data. The maximum likelihood estimate of the corresponding transition probability $q_{i_\ell \cdots i_0}$ is then

$$\hat{q}_{i_\ell \cdots i_0} = \frac{n_{i_\ell \cdots i_0}}{n_{i_\ell \cdots i_1 +}},$$

where

$$n_{i_\ell \cdots i_1 +} = \sum_{i_0=1}^{m} n_{i_\ell \cdots i_0}$$

and the log-likelihood of the entire sequence of observations is written

$$LL = \sum_{i_\ell, \ldots, i_0 = 1}^{m} n_{i_\ell \cdots i_0} \log(\hat{q}_{i_\ell \cdots i_0}).$$

1.2 The Need for Parsimonious Models of High-Order Markov Chains

Markov chains are well suited for the representation of high-order dependencies between successive observations of a random variable. Unfortunately, as the order ℓ of the chain and the number m of possible values increase, the number of independent parameters increases exponentially and rapidly becomes too large to be estimated efficiently, or even identifiably, with data sets of the sizes typically encountered in practice. Table 1 gives the number of independent parameters for different combinations of ℓ and m.

The mixture transition distribution model was introduced to approximate high-order Markov chains with far fewer parameters than the fully parameterized model. Each element of a transition matrix is the probability of observing an event at time t given the events observed at times $(t-\ell)$ to $(t-1)$. In the MTD model, the effect of each lag upon the present is considered separately and the conditional probability is modeled by

$$P(X_t = i_0 | X_{t-\ell} = i_\ell, \ldots, X_{t-1} = i_1) = \sum_{g=1}^{\ell} \lambda_g q_{i_g i_0},$$

where the $q_{i_g i_0}$ are the probabilities of an $m \times m$ transition matrix and λ_g is the weight parameter associated with lag g. This model has only $m(m-1) + (\ell - 1)$ independent parameters and each additional lag adds only one additional parameter. Table 1 shows that when the order is greater than 1, the MTD model is far more parsimonious than the corresponding fully parameterized Markov chain. Therefore, it can be used to estimate high-order transition matrices, even when the amount of data is relatively small.

Parsimonious modeling can also make interpretation easier. A high-order Markov chain can have hundreds or thousands of parameters and it can be difficult to interpret the estimates. On the other hand, a MTD model is generally composed of only one small transition matrix and a vector of lag parameters which are easier to interpret.

1.3 Two Examples

The following two examples show situations where a high-order relationship between successive events cannot be efficiently represented by a standard Markov chain. On the other hand, a MTD model is appropriate.

TABLE 1
Maximal number of independent parameters for different Markov chains and MTD models

Number of values m	Order ℓ	Markov chain	MTD model
2	1	2	2
	2	4	3
	3	8	4
	4	16	5
3	1	6	6
	2	18	7
	3	54	8
	4	162	9
5	1	20	20
	2	100	21
	3	500	22
	4	2,500	23
10	1	90	90
	2	900	91
	3	9,000	92
	4	90,000	93

To evaluate the quality of fit of a model and to compare different models, we use the Bayesian information criterion (BIC). This statistic is defined by

$$BIC = -2LL + p\log(n), \tag{6}$$

where LL is the log-likelihood of the model, p is the number of independent parameters and n is the number of components in the log-likelihood. The model achieving the lowest BIC is chosen. The difference between the BIC values for different models is an approximation to twice the logarithm of a Bayes factor for one model against the other (Schwarz, 1978; Kass and Raftery, 1995). Katz (1981) discussed the use of BIC for Markov chains. Bayes factors have the advantage over standard significance tests of being validly defined for the comparison of nonnested models and also for the comparison of multiple models, both of which we have to deal with here. We do not take into account the parameters estimated to be zero; this is indicated by the derivation of the BIC approximation and is also the convention in counting degrees of freedom for models for categorical data. The number of parameters for a given model can thus be lower than the number given in Table 1.

In the first example, we study a series of wind directions at Koeberg, South Africa. We have a data set of size 744 giving the hourly wind directions during the month of May 1985. The original data appeared in MacDonald and Zucchini (1997), where they were coded into 16 directions. For the purpose of our example, we have recoded the data into $m = 4$ directions: N, NNE, NE and ENE were recoded as state 1, E, ESE, SE and SSE were recoded as state 2 and so on. To have the same number of components in the log-likelihood of each model (730), we conditioned on the first 14 data values and so did not include their contributions to the log-likelihood. Table 2 summarizes our results. The independence model is worse than any Markovian model, according to the BIC criterion, showing that there is dependence between successive observations. Among the fully parameterized Markov chains, the best result is achieved by the first-order model whose *BIC* is equal to 899.1. By examining the reduced form, R_2, of the transition matrix of the second-order Markov chain, we can see that some rows are very poorly estimated:

$R_2 =$

X_{t-2}	X_{t-1}	X_t = 1	2	3	4	Number of data
1	1	0.8482	0.0625	0.0089	0.080	112
2	1	0.6667	0.3333	0	0	15
3	1	0	0.5000	0	0.5000	2
4	1	0.6364	0.0909	0	0.2727	11
1	2	0.3571	0.3571	0.1429	0.1429	14
2	2	0.0365	0.9051	0.0584	0	274
3	2	0.0370	0.7778	0.1852	0	27
4	2	0	0	0	0	0
1	3	0	0	1	0	1
2	3	0	0.4348	0.5217	0.0435	23
3	3	0.0180	0.1441	0.8198	0.0181	111
4	3	0	0.1250	0.7500	0.1250	8
1	4	0.0769	0	0	0.9231	13
2	4	0	0	0.5000	0.5000	2
3	4	0.2500	0	0.2500	0.5000	4
4	4	0.0796	0	0.0531	0.8673	113

For instance, the third row was estimated with only two observations, which is obviously not enough. This suggests that a more parsimonious model of the transition matrix may be useful.

Table 2 shows that the MTD model is preferred to the first-order Markov chain by the BIC criterion. The best model is of order 2 and has a BIC value of 859.3. The MTD 2 model has parameters $\lambda_1 = 0.7569$ for the first lag, $\lambda_2 = 0.2431$ for the second and the transition matrix is

$$Q = \begin{bmatrix} 0.8301 & 0.0689 & 0.0077 & 0.0933 \\ 0.0369 & 0.9012 & 0.0619 & 0 \\ 0.0155 & 0.1553 & 0.8070 & 0.0222 \\ 0.0779 & 0 & 0.0528 & 0.8693 \end{bmatrix}.$$

TABLE 2
Modeling of the wind direction at Koeberg in May 1985

Model	Number of parameters	LL	BIC
Independence	3	−954.8	1929.4
MC 1	11	−413.3	899.1
MC 2	27	−374.9	927.9
MC 3	39	−346.2	949.5
MTD 2	**11**	**−393.4**	**859.3**
MTD 3	12	−393.2	865.6

NOTE: MC stands for Markov chain, MTD stands for mixture transition distribution model and the number following the name of a model is its order. The model with the best BIC value is shown in bold.

The reduced form $\hat{R}_2$ of the transition matrix obtained through the MTD 2 model is

$$\hat{R}_2 = \begin{array}{cc|cccc} & & \multicolumn{4}{c}{X_t} \\ X_{t-2} & X_{t-1} & 1 & 2 & 3 & 4 \\ 1 & 1 & 0.8301 & 0.0689 & 0.0077 & 0.0933 \\ 2 & 1 & 0.6373 & 0.2713 & 0.0209 & 0.0705 \\ 3 & 1 & 0.6321 & 0.0899 & 0.2020 & 0.0760 \\ 4 & 1 & 0.6472 & 0.0522 & 0.0187 & 0.2819 \\ 1 & 2 & 0.2297 & 0.6989 & 0.0487 & 0.0227 \\ 2 & 2 & 0.0369 & 0.9012 & 0.0619 & 0 \\ 3 & 2 & 0.0317 & 0.7199 & 0.2430 & 0.0054 \\ 4 & 2 & 0.0469 & 0.6821 & 0.0597 & 0.2113 \\ 1 & 3 & 0.2135 & 0.1343 & 0.6127 & 0.0395 \\ 2 & 3 & 0.0207 & 0.3366 & 0.6258 & 0.0169 \\ 3 & 3 & 0.0155 & 0.1553 & 0.8070 & 0.0222 \\ 4 & 3 & 0.0306 & 0.1175 & 0.6236 & 0.2283 \\ 1 & 4 & 0.2607 & 0.0168 & 0.0418 & 0.6807 \\ 2 & 4 & 0.0679 & 0.2191 & 0.0550 & 0.6580 \\ 3 & 4 & 0.0627 & 0.0377 & 0.2361 & 0.6635 \\ 4 & 4 & 0.0779 & 0 & 0.0528 & 0.8693 \end{array}.$$

It can be seen that the rows of R_2 that were estimated with a large number of data points have almost identical fitted values in the full Markov chain, R_2, and in the MTD model, $\hat{R}_2$. On the other hand, rows estimated with a very small number of data points in R_2 can be different and tend to be "smoother" in $\hat{R}_2$. This example suggests that the MTD model can be used to improve the estimation of a Markov chain by a kind of smoothing of the raw empirical transition probabilities without too much modification of what is already well estimated.

As another example, MacDonald and Zucchini (1997) presented a time series of the daily counts of epileptic seizures of a patient. We consider here a binary variable describing for each day whether the patient had no epileptic seizure or at least one. Here, 0 denotes no epileptic seizure and 1 denotes at least one. The data are to be read row by row, from left to right:

```
0100001  1011110  0111111
0110111  1111101  0111110
0011101  0101000  0000010
0000100  0100010  0010010
0110110  0011011  1111101
1111111  0111011  0011101
0010100  0100100  0000011
0000010  1110000  0101110
0000011  0000000  0000000
0000000  0000000  0
```

The complete time series is of length 204, but as before we conditioned on the first 14 observations when computing the log-likelihood. We estimated Markov chains as well as MTD models and the results are shown in Table 3.

In this example, the results achieved by the MTD model show that the present is best explained by using as many as eight lags. Computing the fully parameterized eighth-order Markov chain would not have been realistic here because of the huge number of parameters involved. The parameters of the best MTD model are

$$\begin{array}{lll} \lambda_1 = 0.1278, & \lambda_2 = 0.0963, & \lambda_3 = 0.0953, \\ \lambda_4 = 0.2605, & \lambda_5 = -0.1778, & \lambda_6 = 0.0932, \\ \lambda_7 = 0.1389, & \lambda_8 = 0.3658 & \end{array}$$

TABLE 3
Modeling of the epileptic data

Model	Number of parameters	LL	BIC
Independence	1	−129.3	263.9
MC 1	2	−122.6	255.6
MC 2	4	−119.3	259.6
MC 3	8	−117.2	276.3
MC 4	16	−111.5	306.9
MC 5	27	−101.9	345.4
MTD 2	3	−119.5	254.7
MTD 3	4	−117.7	256.4
MTD 4	5	−113.2	252.7
MTD 5	6	−112.4	256.2
MTD 6	7	−110.4	257.6
MTD 7	8	−107.9	257.7
MTD 8	**9**	**−102.3**	**251.9**
MTD 9	10	−100.5	253.5
MTD 10	11	−100.0	257.7

NOTE: MC stands for Markov chain, MTD stands for mixture transition distribution model and the number following the name of a model is its order. The model with the best BIC value is shown in bold.

for the lag parameters and

$$Q = \begin{pmatrix} 0.8750 & 0.1250 \\ 0.1723 & 0.8277 \end{pmatrix}$$

for the transition matrix.

The most important lag is the eighth, which explains why this model does better than shorter ones. Adding a ninth or a tenth lag does not improve the results. The fifth lag is negative, so the relationship between this lag and the present is inverse compared to the influence of the other lags. Since the greater probabilities of Q are on the first diagonal, there is generally a positive association between a lagged value and the present. If, for instance, the patient had an epileptic seizure at time $(t-4)$, this increases the probability that he will have an epileptic seizure at time t.

2. THE MTD MODEL

In this section and the following ones we go into more detail about the MTD model. The MTD model was introduced by Raftery (1985a) for the modeling of time-homogeneous high-order Markov chains. In this section we define the model, we give a limit theorem and we provide some results about the correlation structure. This section is based essentially on Raftery (1985a, b), Adke and Deshmukh (1988), Haney (1993) and Raftery and Tavaré (1994).

2.1 Model

Let $\{X_t\}$ be a sequence of random variables taking values in the finite set $N = \{1, \ldots, m\}$. In an ℓth-order Markov chain, the probability that $X_t = i_0$, $i_0 \in N$, depends on the combination of values taken by $X_{t-\ell}, \ldots, X_{t-1}$. In the MTD model, the contributions of the different lags are combined additively. Then

$$\begin{aligned} &P(X_t = i_0 | X_{t-\ell} = i_\ell, \ldots, X_{t-1} = i_1) \\ &\quad = \sum_{g=1}^{\ell} \lambda_g P(X_t = i_0 | X_{t-g} = i_g) \qquad (7) \\ &\quad = \sum_{g=1}^{\ell} \lambda_g q_{i_g i_0}, \end{aligned}$$

where $i_\ell, \ldots, i_0 \in N$, the probabilities $q_{i_g i_0}$ are elements of an $m \times m$ transition matrix $Q = [q_{i_g i_0}]$, each row of which is a probability distribution (i.e., each row sums to 1 and the elements are nonnegative) and $\lambda = (\lambda_\ell, \ldots, \lambda_1)'$ is a vector of lag parameters. Note that we adopt the convention that each *row* of the transition matrix Q is a probability distribution, whereas in some papers it is each *column* of Q that is taken to be a probability distribution.

To ensure that the results of the model are probabilities, that is,

$$0 \le \sum_{g=1}^{\ell} \lambda_g q_{i_g i_0} \le 1, \qquad (8)$$

the vector λ is subject to the constraints

$$\sum_{g=1}^{\ell} \lambda_g = 1, \qquad (9)$$

$$\lambda_g \ge 0. \qquad (10)$$

Raftery and Tavaré (1994) showed that the constraints (10) can be removed, but that the model can then produce results which are no longer probabilities. It is then necessary, so that the conditional probabilities in (7) are well defined, to impose the new set of constraints

$$T q_i^- + (1-T) q_i^+ \ge 0 \quad \forall i \in \{1, \ldots, m\}, \qquad (11)$$

where

$$T = \sum_{\substack{g=1 \\ \lambda_g \ge 0}}^{m} \lambda_g,$$

$$q_i^- = \min_{1 \ge j \ge m} q_{ij},$$

$$q_i^+ = \max_{1 \ge j \ge m} q_{ij}.$$

Equation (7) gives the probability corresponding to each combination of $i_\ell, \ldots, i_0$ individually. The model can also be written in matrix form, giving the whole distribution of X_t. Let χ_t and $\hat{\chi}_t$ be the vectors defined by (1) and (2). The MTD model can then be rewritten as

$$\hat{\chi}_t' = \sum_{g=1}^{\ell} \lambda_g \chi_{t-g}' Q. \qquad (12)$$

Since each row of the transition matrix Q is a probability distribution and therefore sums to 1, this matrix has $m(m-1)$ independent parameters. In addition, an ℓth-order model has ℓ lag parameters $\lambda_\ell, \ldots, \lambda_1$, but by (9) only $(\ell - 1)$ of them are independent. An ℓth-order MTD model thus has $m(m-1) + (\ell - 1)$ independent parameters, which is far more parsimonious than the corresponding fully parameterized Markov chain (see Table 1). Moreover, each additional lag adds only one parameter to the model.

2.2 Limiting Behavior of the MTD Model

The MTD model has the same equilibrium distribution as the first-order Markov chain with transition matrix Q, no matter what the MTD order is. This is somewhat appealing, because it means that the parameters defining the equilibrium or marginal distribution are specified separately from those defining the lag distribution of the model.

Let $\pi = (\pi_1, \ldots, \pi_m)'$ be the limit distribution of the first-order Markov chain, that is, the long-term distribution of values $1, \ldots, m$ given the first-order model. Raftery (1985a) gave a theorem giving conditions ensuring that the MTD model has π as limit distribution. Adke and Deshmukh (1988) showed that these conditions are stronger than needed and they gave a more general theorem stating that if the transition matrix Q of the MTD model admits a limit probability distribution ω, that is, if there is a vector ω such that

$$\lim_{n\to\infty} Q^n = \iota \otimes \omega',$$

where ι is a vector of 1's of size $m \times 1$, $\omega = (\omega_1, \ldots, \omega_m)'$, $\omega_g \geq 0$, $\sum_{g=1}^m \omega_g = 1$ and $\otimes$ is the Kronecker product, then the MTD model has the same limit behavior as the fully parameterized high-order Markov chain. The same result holds when the MTD model is defined on an arbitrary state space, not necessarily finite or even countable.

2.3 Autocorrelation Structure

Since the MTD model is similar to an autoregressive (AR) model, it is interesting to verify whether it has the same type of autocorrelation structure. Let $\{X_t, t \in \mathbb{Z}\}$ be an ℓth-order autoregressive process defined by

$$X_t = \sum_{g=1}^{\ell} \lambda_g X_{t-g} + \varepsilon_t,$$

where λ_g are real parameters and the ε_t form a set of noncorrelated random variables with expectation zero and variance σ^2.

The order k autocorrelation function of the process, denoted by ρ_k, is the correlation between X_t and X_{t+k}. These autocorrelations satisfy the Yule–Walker equations, namely

$$\rho_k = \sum_{g=1}^{\ell} \lambda_g \rho_{k-g}. \tag{13}$$

The autocorrelation structure of the MTD model does not satisfy (13), but Raftery (1985a) showed that the entire system of bivariate distributions does satisfy a system of linear equations similar to the Yule–Walker equations. Let $B(k)$ be an $m \times m$ matrix whose elements are

$$b_{ij}(k) = P(X_t = i, X_{t+k} = j), \qquad i, j = 1, \ldots, m,\ k \in \mathbb{Z}^*,$$

and $B(0) = \operatorname{diag}(\pi_1, \ldots, \pi_m)$, where $(\pi_1, \ldots, \pi_m)$ is the first-order limit distribution of the process. Then the bivariate distributions $B(k)$ satisfy the system of matrix equations:

$$B(k) = \sum_{g=1}^{\ell} \lambda_g Q B(k-g). \tag{14}$$

Theorem 3 of Raftery (1985a) states that these equations have a unique solution in the following cases:

1. If $\ell = 2$, (14) has a unique solution if $0 < \lambda_1 \leq 1$.
2. If $\ell = 3$, (14) has a unique solution if $\lambda_g \geq 0$, $g = 1, 2, 3$, and if Q has at least one column whose elements are all strictly greater than zero.
3. If $\ell \geq 4$, (14) has a unique solution if $\lambda_g \geq 0$, $g = 1, \ldots, \ell$, and if

$$(1 - \eta_1 - \cdots - \eta_m)(2 - \lambda_1 - \lambda_{\ell-1} - \lambda_\ell) < 1,$$

where η_i is the minimum of the ith column of Q.

Now, suppose that Y_t is a random variable with distribution $Q\chi_t$, so that $P(Y_t = j | X_t = i) = q_{ij}$. Let ρ_k denote the correlation between X_t and X_{t+k}, and let $\tilde{\rho}_k$ be the correlation between X_t and Y_{t+k}. Then the autocorrelations satisfy a system of equations similar to the Yule–Walker equations, namely

$$\rho_k = \sum_{g=1}^{\ell} \lambda_g \tilde{\rho}_{k-g}. \tag{15}$$

It is of particular interest to investigate the range of autocorrelations which can be represented by the MTD model. This question has been addressed by Raftery (1985a, b) and Haney (1993). Raftery studied the autocorrelation structure of the second-order model with $m = 3$ states. He considered the special case where the marginal probabilities are equal and Q is constructed such that it satisfies a set of Yule–Walker equations,

$$Q = \tfrac{1}{3}(1 - |\alpha|)J + \begin{cases} \alpha I, & 0 \leq \alpha \leq 1, \\ |\alpha| E, & -1 \leq \alpha < 0, \end{cases} \tag{16}$$

where $|\alpha| \leq 1$, J is a 3×3 matrix of 1's, I is the 3×3 identity matrix and

$$E = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Equation (15) implies that

$$\begin{aligned} \rho_1 &= \Lambda_1 + \Lambda_2 \rho_1, \\ \rho_2 &= \Lambda_1 \rho_1 + \Lambda_2, \end{aligned} \tag{17}$$

where $\Lambda_1 = \lambda_1 \alpha$ and $\Lambda_2 = \lambda_2 \alpha$. By combining (17) with the set of constraints (8), we find that the range of possible autocorrelations is given by

$$\begin{aligned} \rho_1 &\geq -\tfrac{1}{2}, \\ \rho_1 + \rho_2 &\geq 0, \\ \rho_2 &\geq \{\rho_1(1 + 3\rho_1) - 1\}/(2 + \rho_1) \end{aligned}$$

for $\alpha \geq 0$ and by

$$\begin{aligned} -(1 + 2\rho_1) &\leq \rho_2 \leq -\rho_1, \\ \rho_2(1 + 2\rho_1) &\geq 2\rho_1(1 + \rho_1) - 1, \\ \rho_2 - 1 &\leq (\rho_1 + 1)(\rho_1 - \rho_2) \end{aligned}$$

for $\alpha < 0$. Figure 1 shows the possible range of autocorrelation of the MTD model. In spite of the large number of constraints, this range is almost as great as for the standard AR(2) model.

Haney (1993) investigated the case $\ell = 3$ and $m = 2$. Let $P^k_{i_0 i_k}$ denote the bivariate probability of being in state i_0 at time $t = 0$ and in state i_k at time $t = k$. By means of numerical simulations, Haney computed the probabilities P^k_{11} and P^k_{22} for $k = \{1, 5\}$. After one period ($k = 1$), the MTD model can represent roughly the same range of situations as the fully parameterized Markov chain. On the other hand, after $k = 5$ periods the restrictions imposed by the MTD model appear clearly. In the long term, the MTD model does not allow the representation of all the situations a real Markov chain does, but this is not surprising given the number of constraints that are imposed. In particular, the MTD model finds it hard to represent situations where P^k_{11} and P^k_{22} are both small.

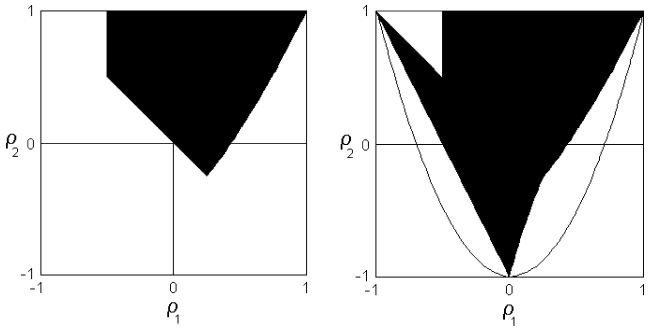

FIG. 1. *Range of possible autocorrelations for the second-order three-state MTD model. The case $\alpha \geq 0$ is represented on the left and the general case is on the right. The continuous line on the right indicates the limit of the autocorrelation range for the standard* AR(2) *model.* Source: *Raftery* (1985a, b).

3. GENERALIZATIONS OF THE MTD MODEL

In this section, we present several generalizations of the MTD model that allow it to fit data more accurately and to be applied to different fields.

3.1 The Multimatrix MTD Model

In the original MTD model, the same transition matrix Q is used to represent the relationship between each lag and the present. Raftery (1985b) proposed relaxing this assumption by using a different $m \times m$ transition matrix for each lag. Berchtold (1995, 1996, 1998) developed and generalized this idea.

A first generalization along these lines is called MTDg and is defined as

$$\begin{aligned} &P(X_t = i_0 | X_{t-\ell} = i_\ell, \ldots, X_{t-1} = i_1) \\ &\quad = \sum_{g=1}^{\ell} \lambda_g q^{(g)}_{i_g i_0} \end{aligned} \tag{18}$$

or, for the complete distribution at time t,

$$\hat{\chi}'_t = \sum_{g=1}^{\ell} \lambda_g \chi'_{t-g} Q_g,$$

where $\lambda_\ell, \ldots, \lambda_1$ are the lag parameters and $Q_g = [q^{(g)}_{i_g i_0}]$ is an $m \times m$ transition matrix giving the relationship between the gth lag and the present. This model has $\ell m(m-1) + (\ell - 1)$ independent parameters. However, as pointed out by Raftery (1985b), (18) can be written more generally as

$$\begin{aligned} &q_{i_\ell, \ldots, i_1, i_0} \\ &\quad = \begin{cases} \displaystyle\sum_{g=1}^{\ell} a^{(g)}_{i_g i_0}, & i_0 = 1, \ldots, m-1, \\ \displaystyle 1 - \sum_{k=1}^{m-1} q_{i_\ell, \ldots, i_1, k}, & i_0 = m, \end{cases} \end{aligned} \tag{19}$$

where

$$q_{i_\ell, \ldots, i_1, i_0} = P(X_t = i_0 | X_{t-\ell} = i_\ell, \ldots, X_{t-1} = i_1)$$

and $a^{(g)}_{i_g i_0} = \lambda_g q^{(g)}_{i_g i_0}$. Using the form (19), the MTDg model has only $\ell m(m-1)$ independent parameters. Even though it is less parsimonious than the original MTD model, it has proved to be of use in different situations.

More generally, the MTD modeling principle is not restricted to the use of transition matrices giving the relationship between *one* lagged period and the present. Let $\mathcal{Q}$ be the set of all transition matrices which can be built between a subset of size strictly smaller than ℓ of the lagged periods $t-\ell, \ldots, t-1$ and the period t. For instance, if $\ell = 3$,

$$\mathcal{Q} = \left\{ \begin{array}{l} Q_1 = \big[P(X_t = i_0 | X_{t-1} = i_1)\big] \\ Q_2 = \big[P(X_t = i_0 | X_{t-2} = i_2)\big] \\ Q_3 = \big[P(X_t = i_0 | X_{t-3} = i_3)\big] \\ Q_4 = \big[P(X_t = i_0 | X_{t-2} = i_2, X_{t-1} = i_1)\big] \\ Q_5 = \big[P(X_t = i_0 | X_{t-3} = i_3, X_{t-1} = i_1)\big] \\ Q_6 = \big[P(X_t = i_0 | X_{t-3} = i_3, X_{t-2} = i_2)\big] \end{array} \right\}.$$

Note that the definition of $\mathcal{Q}$ does not include the matrix using all lagged periods $t-\ell, \ldots, t-1$ in Q, since this is the matrix we want to model.

By using all the matrices of the set $\mathcal{Q}$, we achieve a more general family of MTD models. The complete model is generally overparameterized and in practice only a subset of the members of Q would be used. For instance, in the case $\ell = 3$, the MTDg model defined above uses only the matrices Q_1, Q_2 and Q_3. As shown by Berchtold (1998), this type of modeling still has the same limit distribution as the real first-order Markov chain.

3.2 Infinite-Lag MTD Models

In the Markov chain framework, the number of lagged periods used in specifying the conditional distribution of the present state is finite. However, in some situations it can be useful to use all the past values. Even if this is no longer a Markov chain, the MTD model can be adapted to this case. This idea was first considered by Mehran (1989a, b) for discrete-valued time series and then developed by Le, Martin and Raftery (1996) in a more general context. Since $\ell = \infty$, the lag parameters have to be reparameterized, and it is hard to estimate the parameters λ and Q. Mehran proposed two solutions involving parameterizing the λ_g sequence, namely

$$\begin{aligned} &\lambda_g = \xi^{g-1}(1-\xi), && 0 < \xi < 1, \\ &\lambda_g = g^{-\alpha} - (g+1)^{-\alpha}, && \alpha > 0. \end{aligned}$$

These two formulations lead to time-decreasing lag parameters and they sum to 1. Another formulation is due to Le, Martin and Raftery (1996):

$$\begin{aligned} \lambda_g &= \frac{d(1-d)\cdots(g-1-d)}{g!} \\ &= \frac{(g+d-1)!}{g!(d-1)!}. \end{aligned}$$

These weights are similar to the partial linear autoregression coefficients for the fractionally differenced ARIMA$(0, d, 0)$ process [Hosking (1981)].

Parameterizing the λ_g's leads to a more parsimonious model. On the other hand, we must be aware that in some situations the lag parameters are not strictly decreasing in time. The epileptic seizure data in Section 1 provide an example of this.

3.3 Missing Data and Finite Length Sequences

In practice all sequences of data are of finite length, which seems to prevent the use of the infinite-lag models of Section 3.2. Moreover, some data sets can have missing data. An example of such data is the rotation sampling scheme used in the U.S. Current Population Survey. Each unit is surveyed during a first period of 4 consecutive months, then dropped from the sample for the next 8 months and surveyed again during a final period of 4 months. This creates a sequence with a gap of 8 months.

Consider the time series

$$X_0\ X_1\ \ldots\ X_{t-k-1}\ ?\ X_{t-k+1}\ \ldots\ X_t\ \ldots,$$

where X_{t-k} is missing. The probability $q_{i_k i_0}$ appearing in the MTD model has to be estimated and one possible choice is to once again use a MTD model. An estimate $\hat{q}_{i_k i_0}$ of $q_{i_k i_0}$ is given by

$$\hat{q}_{i_k i_0} = \sum_{r=1}^{\infty} \lambda_r q^{(2)}_{i_{k+r} i_0}, \tag{20}$$

where $q^{(2)}_{i_{k+r} i_0}$ denotes the two-step transition between i_{k+r} and i_0 in the matrix Q^2. The same relationship can be used for finite lag models by replacing the infinite sum by a sum up to ℓ.

When a term $q^{(2)}_{i_{k+r} i_0}$ of (20) is itself missing, it may be replaced by

$$\hat{q}^{(2)}_{i_{k+r} i_0} = \sum_{s=1}^{\infty} \lambda_s q^{(3)}_{i_{k+r+s} i_0},$$

where $q^{(3)}_{i_{k+r+s}i_0}$ denotes the three-step transition between i_{k+r+s} and i_0 in the matrix Q^3. The same procedure can be generalized to any number of missing data values.

Berchtold (1996) derived the same type of relationship for the MTDg model. In this case, (20) is replaced by

$$\hat{q}_{i_k i_0} = \sum_{r=1}^{\infty} \lambda_r q^{(k+r)}_{i_{k+r} i_0},$$

where $q^{(k+r)}_{i_{k+r}i_0}$ denotes the transition between i_{k+r} and i_0 in the matrix Q_{k+r}.

A time series of finite length can be viewed as an infinite sequence that has data missing from a certain point onward. Using (20), the infinite-lag MTD model becomes

$$\begin{aligned} &P(X_t = i_0 | X_{t-1} = i_1, X_{t-2} = i_2, \ldots) \\ &\quad = \sum_{g=1}^{\infty} \lambda_g q_{i_g i_0} \\ &\quad = \sum_{g=1}^{t-1} \lambda_g q_{i_g i_0} + \theta_{t-1} \pi_{i_0}, \end{aligned}$$

where $\theta_{t-1} = 1 - \sum_{g=1}^{t-1} \lambda_g$ and π_{i_0} is the probability of i_0 in the limit distribution of the transition matrix Q.

Here we have mentioned possible MTD-specific ways of dealing with missing data and infinite-lag models. However, more general prescriptions for missing data, such as those surveyed by Schafer (1997), could also be applied to MTD models, although they might be more complex to implement.

3.4 Infinite Denumerable State Spaces

The MTD model was initially designed for finite state spaces. However, discrete-valued time series can often take an infinite but countable number of different values. Counts of events in a point process are an important example. We can think, for instance, of the number of flights landing each hour at an airport or of the number of car crashes in a city each day.

The main problem in applying the MTD model with an infinite state space is that the transition matrix Q is also of infinite size. It must then be respecified with a finite number of parameters. Following Raftery (1985b), a simple method is to consider a random vector (Y, Z) having the desired marginal distribution and to define $q_{zy} = P(Y = y | Z = z)$. A first possibility is the bivariate Poisson distribution of Holgate (1964). Let Y and Z be two Poisson variables with the same mean μ. Then Holgate's bivariate Poisson distribution is defined by

$$\begin{aligned} &P(Y = y, Z = z) \\ &\quad = e^{-(2\mu - \zeta)} \sum_{h=0}^{\min\{y,z\}} \frac{(\mu - \zeta)^{y+z-2h} \zeta^h}{(y-h)!(z-h)!h!}, \end{aligned}$$

where ζ is the covariance of Y and Z. Since

$$P(Z = z) = \frac{e^{-\mu} \mu^z}{z!},$$

the elements of the transition matrix Q are

$$\begin{aligned} &P(Y = y | Z = z) \\ &\quad = \frac{e^{-(\mu-\zeta)} z!}{\mu^z} \sum_{h=0}^{\min\{y,z\}} \frac{(\mu - \zeta)^{y+z-2h} \zeta^h}{(y-h)!(z-h)!h!}. \end{aligned}$$

Raftery (1985b) also proposed the bivariate negative binomial model of Johnson and Kotz (1969) as a second possibility. In this case, the elements of the transition matrix Q are

$$P(Y = y | Z = z) = \binom{y + z + \nu - 1}{y} p^y (1-p)^{\nu+z},$$

where $\nu > 0$ and $0 \le p \le 1$.

3.5 General State Spaces

The MTD model was proposed as an approximation of high-order Markov chains with a finite number of states, but it can be easily extended to the modeling of more general processes, with an arbitrary state space. This approach was proposed in several papers including Martin and Raftery (1987), Adke and Deshmukh (1988), Raftery (1993), Le, Martin and Raftery (1996) and Wong and Li (2000). It provides a powerful tool for the analysis of non-Gaussian time series and can represent such nonlinear or non-Gaussian features as outliers, change points, bursts of high variance activity or volatility, and flat stretches.

Equation (7) can be generalized as follows. Let $\{X_t; t \in \mathbb{N}\}$ be a sequence of random variables taking values in an arbitrary state space. Let $X_0^t = (X_0, \ldots, X_t)$. Then

$$F(x_t | x_0^{t-1}) = \sum_{g=1}^{\ell} \lambda_g G_g(x_t | x_{t-g}), \tag{21}$$

where $F(x_t | x_0^{t-1})$ is the conditional cumulative distribution function of X_t given its past, $G_g(x_t | x_{t-g})$ is a conditional cumulative distribution function of X_t

given the gth lag, $\sum_{g=1}^{\ell} \lambda_g = 1$ and $\lambda_g \geq 0$ for $g = 1, \ldots, \ell$. The nonnegativity constraints are needed to ensure that all the probability densities specified by the model are positive.

Le, Martin and Raftery (1996) proposed specifying the G_g to be Gaussian, namely

$$G_g(x_t|x_{t-g}) = \Phi\left(\frac{x_t - \phi_g x_{t-g}}{\sigma_g}\right).$$

They noted that even though the conditional distributions are mixtures of Gaussian distributions, the resulting model is able to represent non-Gaussian behavior.

Le, Martin and Raftery (1996) also proposed two generalizations of this model. First, by including a supplementary term, the standard AR(ℓ) process becomes a particular case of the MTD model:

$$F(x_t|x_0^{t-1}) = \lambda_0 \Phi\left(\frac{x_t - \sum_{g=1}^{\ell} \phi_{0g} x_{t-g}}{\sigma_0}\right) + \sum_{g=1}^{\ell} \lambda_g \Phi\left(\frac{x_t - \phi_g x_{t-g}}{\sigma_g}\right). \tag{22}$$

They then proposed to add a further independent component allowing specific modeling of outliers. Their final model is written

$$F(x_t|x_0^{t-1}) = \lambda_0 \Phi\left(\frac{x_t - \sum_{g=1}^{\ell} \phi_{0g} x_{t-g}}{\sigma_0}\right) + \sum_{g=1}^{\ell} \lambda_g \Phi\left(\frac{x_t - \phi_g x_{t-g}}{\sigma_g}\right) + \lambda_{\ell+1} \Phi\left(\frac{x_t}{\sigma_{\ell+1}}\right), \tag{23}$$

where $\sigma_{\ell+1}$ is large and $\sum_{g=0}^{\ell+1} \lambda_g = 1$. It has only $4\ell + 3$ independent parameters, which is quite parsimonious. Models (21), (22) and (23) are called Gaussian mixture transition distribution models (GMTD). By specifying

$$\phi_1 = \cdots = \phi_\ell = 1,$$

$$\sum_{g=1}^{\ell} \phi_{0g} = 1,$$

a special case of the GMTD model called the random walk GMTD model is defined. It generalizes the usual random walk with only $(3\ell + 2)$ independent parameters.

The GMTD model can represent time series that are well represented by an autoregressive process, but it also has the ability to capture occasional ruptures in the sequence such as bursts, outliers or even flat stretches without explicitly specifying them. Outliers may be captured by a term with a large variance σ_g^2 and a small λ_g. Bursts can be modeled with a large λ_g, and a random walk GMTD is able to represent flat stretches. It is also possible to represent time series that present heteroskedasticity and multimodal marginal distributions. Examples are given in Sections 5.4 and 5.5.

Le, Martin and Raftery (1996) also examined the stationarity and autocorrelation properties of the GMTD model. Their Theorem 1 states that the process X_t given by (22) is first-order stationary if the roots $z_1, \ldots, z_\ell$ of the equation

$$1 - \sum_{g=1}^{\ell} (\lambda_0 \phi_{0g} + \lambda_g \phi_g) z^{-g} = 0$$

all lie inside the unit circle. Their Theorem 2 states that the process X_t given by (22) is second-order stationary if the roots $z_1, \ldots, z_\ell$ of the equation

$$1 - \sum_{g=1}^{\ell} \lambda_g \phi_g^2 z^{-g} = 0$$

all lie inside the unit circle.

When there are only two mixture components, that is, $\ell = 2$ and $\lambda_0 = 0$, the region given by the first-order conditions is given by

$$\lambda_1 \phi_1 + \lambda_2 \phi_2 < 1,$$

$$-\lambda_1 \phi_1 + \lambda_2 \phi_2 < 1,$$

$$-\lambda_2 \phi_2 < 1,$$

and the region corresponding to second-order stationarity is given by

$$\lambda_1 \phi_1^2 + \lambda_2 \phi_2^2 < 1.$$

Figure 2 displays these two regions.

As is the case for the original MTD model, the autocorrelations of the GMTD model satisfy a system of equations similar to the Yule–Walker equations. Let X_t be a second-order stationary process with mean zero. If ρ_k denotes the kth autocorrelation, then

$$E(X_t X_{t-k}) = \sum_{g=1}^{\ell} (\lambda_0 \phi_{0g} + \lambda_g \phi_g) E(X_{t-k} X_{t-g})$$

and since the process is second-order stationary,

$$\rho_k = \sum_{g=1}^{\ell} (\lambda_0 \phi_{0g} + \lambda_g \phi_g) \rho_{|k-g|} \tag{24}$$

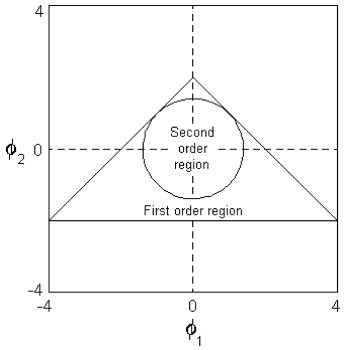

FIG. 2. *Stationarity regions for the second-order GMTD model.* $\lambda_1 = \lambda_2 = 0.5$. Source: *Le, Martin and Raftery* (1996).

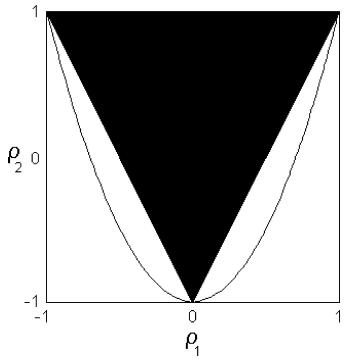

FIG. 3. *Range of possible autocorrelations for the second-order GMTD model. The continuous line indicates the limit of the autocorrelation range for the standard* AR(2) *model.* Source: *Le, Martin and Raftery* (1996).

for $k = 1, \ldots, \ell$. For $\lambda_0 = 0$ and $\ell = 2$, (24) becomes

$$\rho_1 = \lambda_1\phi_1 + \lambda_2\phi_2\rho_1,$$
$$\rho_2 = \lambda_1\phi_1\rho_1 + \lambda_2\phi_2,$$

and Figure 3 presents the admissible regions for the GMTD and the corresponding AR(2) process. It appears that even without the AR term, the range of autocorrelations for the GMTD model is almost as broad as for the standard AR(2) process.

Wong and Li (2000) proposed a further generalization of the GMTD called the mixture autoregressive model (MAR). It is defined by

$$F(x_t|x_0^{t-1}) = \sum_{k=1}^{K} \lambda_k \cdot \Phi\left(\frac{x_t - \phi_{k0} - \phi_{k1}x_{t-1} - \cdots - \phi_{kp_k}x_{t-p_k}}{\sigma_k}\right), \tag{25}$$

where $\sum_{k=1}^{K} \lambda_k = 1$, $\lambda_1, \ldots, \lambda_K > 0$, K is the number of components in the mixture and $p_1, \ldots, p_K$ are the numbers of lags in each component. By constraining some of the ϕ parameters to equal zero, each version of the GMTD, including the random walk GMTD, can be written as a special case of the MAR model. The conditions for first- and second-order stationarity are similar to those given for the GMTD model.

The full MAR model has $3K - 1 + \sum_{k=1}^{K} p_k$ independent parameters, which is much greater than the number required by the GMTD model and can lead to overparameterization problems. Moreover, as noted by Le, Martin and Raftery (1996), there are potential near nonidentifiability problems with this class of models, and these problems can only become worse with the use of a full MAR specification. Nevertheless, the MAR model extends the range of situations which can be modeled by the Gaussian mixture transition distribution method.

3.6 MTD Approximation of Conditional Distributions for Spatial Data

The MTD model as defined by Raftery (1985a) was designed for the modeling of Markov chains, or more generally of time series. However, the same model may also be useful in a spatial context, replacing the temporal reference by a concept of neighborhood. The resulting model does not define a joint probability distribution, but some experience suggests that it may nevertheless be useful for approximating local conditional distributions at one spatial location given neighbors. This possibility was first mentioned by Raftery and Banfield (1991). Berchtold (2001) motivated this concept in the one-dimensional case and gave a practical example in which the MTD model seemed to provide a better local approximation to spatial conditional distributions than the classical Potts model, subject to the caveat above. Other applications appear in Berchtold (1998).

The traditional Markovian process for spatial data is the Markov random field (MRF) model (Besag, 1974). Here we consider the two-dimensional case, but this approach can be applied to any number of dimensions greater than or equal to 1. Let X be a lattice each site of which is labeled with indices (i, j). To each site corresponds a random variable X_{ij} taking values in the finite set $\{1, \ldots, m\}$. Here we consider the problem of estimating the probability of observing a particular value of the variable X_{ij} given the values of the other variables in the lattice. Let X_{ij}^{N} denote a neighborhood of X_{ij}, that is, a subset of the entire set of variables. For instance, X_{ij}^{N} can be the set of the four nearest neighbors of X_{ij}, namely $\{X_{i-1,j}, X_{i,j+1}, X_{i+1,j}, X_{i,j-1}\}$. The

conditional probability of observing $X_{ij} = x_{ij}$ given its neighbors is then

$$P(X_{ij} = x_{ik} | X \setminus \{X_{ij}\}) = P(x_{ij} | x_{i-1,j}, x_{i,j+1}, x_{i+1,j}, x_{i,j-1}). \tag{26}$$

As in the case of a high-order Markov chain, the set of explanatory variables (the neighbors) is considered as a whole, but if we make the approximating assumption that the effects of the neighbors upon X_{ij} combine additively, the MTD principle can be used to approximate (26). For simplicity, we use the notation $X_0 = X_{ij}$ and we number the neighbors from X_1 to X_4. Then, applying (7), we write

$$P(x_0|x_1, x_2, x_3, x_4) = \sum_{g=1}^{4} \lambda_g P(x_0|x_g) = \sum_{g=1}^{4} \lambda_g q_{x_g x_0}, \tag{27}$$

where $q_{x_g x_0}$ are the elements of an $m \times m$ transition matrix Q representing the conditional relationship between X_0 and each of its neighbors. The same principle can be extended to different sets of neighbors. It is also possible to use a different transition matrix to model the relationship between X_0 and each of its neighbors.

As outlined by Raftery and Banfield (1991), the main difficulty with the use of the MTD model in the spatial context is that it does not satisfy the Hammersley–Clifford theorem [Grimmett (1973), Besag (1974)], since (27) does not define a correct joint probability over the whole set of random variables. Nevertheless, this method may provide a reasonable approximation to the local structure of the joint distribution, a point also made by Besag, York and Mollié (1991). Note that the spatial MTD model is a dependency network, as defined by Heckerman et al. (2000).

It is interesting to compare the spatial MTD model with the classical Potts model (Domb and Potts, 1951; Kinderman and Snell 1980; Baxter, 1982). In the Potts model, the probability of observing a given realization of the variable X_0 is a function of the *number* of neighbors taking the same value. For instance, the Potts model corresponding to (27) is

$$P(x_0|x_1, x_2, x_3, x_4) = \frac{\exp\{-\beta N_0\}}{\sum_{j=1}^{m} \exp\{-\beta N_j\}},$$

where N_0 denotes the number of neighbors taking value x_0, and N_j, $j = 1, \ldots, m$, is the number of neighbors taking value j. One way of estimating this model is to define $(m-1)$ logistic regressions having the same parameter β,

$$\pi_j = \log\left(\frac{P(X_0 = j|X_1, X_2, X_3, X_4)}{P(X_0 = m|X_1, X_2, X_3, X_4)}\right) = \beta(N_m - N_j), \quad j = 1, \ldots, m-1,$$

and to compute the corresponding maximum pseudo-likelihood estimate of β (Besag, 1975, 1977).

The Potts model can be viewed as involving two steps. First, each neighbor X_k is recoded as a binary variable:

$$\delta_k = \begin{cases} 1, & \text{if } x_k = x_0, \\ 0, & \text{otherwise.} \end{cases}$$

Then, it is supposed that the relative position of each neighbor to the variable X_0 is of no importance and that all that matters is the sum $N_0 = \sum_{k=1}^{4} \delta_k$.

Compared to the MTD model, these two steps are two more levels of simplification which may lead to a loss of flexibility in modeling. So if the original variables are already binary and if it is possible to justify the omission of the relative position of each neighbor, the Potts model may perform well because of its parsimony. On the other hand, when these assumptions are too strong, the MTD model may be an interesting alternative. Berchtold (2001) applied the spatial MTD model to the analysis of a one-dimensional DNA sequence. Each base was modeled using information about both the left and right bases, and the resulting model proved to represent the local conditional distributions better than either the traditional Markov chain or the Potts model.

3.7 MTD Regression Models

It is possible to use the MTD model as the basis for an alternative specification of regression models [Berchtold (1998)]. Suppose the response variable is Y and the independent variables are X_1 and X_2. Then an MTD-like model can be written as

$$P(Y = y|X_1 = x_1, X_2 = x_2) = \lambda_1 P(y|x_1) + \lambda_2 P(y|x_2).$$

Although we have not investigated this in depth, one could imagine several advantages to such a formulation. It provides potentially great flexibility in the specification of the $P(\cdot|\cdot)$ functions, since these are conditional on only one variable, and yet it specifies an entire multivariate response. Also, it can be applied in situations outside standard generalized linear model responses, such as when Y takes values in the simplex, or

a circle, or on a sphere or when Y is a multivariate non-Gaussian response.

A more general formulation is

$$P(Y = y|X_1 = x_1, X_2 = x_2) = \sum_{g=1}^{\ell} \lambda_g P_g(y|x_1, x_2).$$

This would allow a range of generalizations. This is worthy of further investigation.

4. ESTIMATION

Since the introduction of the MTD model in 1985, several estimation methods have been proposed. Most of these methods use algorithms and software that are not broadly available or that can be applied only in special situations. In this section, we review the main estimation methods that have been proposed.

4.1 Numerical Maximization of the Log-Likelihood

The parameters λ and q of the MTD model (7) can be estimated by maximizing the log-likelihood of the model:

$$LL = \sum_{i_\ell,\ldots,i_0=1}^{m} n_{i_\ell,\ldots,i_0} \log\left(\sum_{g=1}^{\ell} \lambda_g q_{i_g i_0}\right), \tag{28}$$

where $n_{i_\ell,\ldots,i_0}$ is the number of sequences of the form

$$X_{t-\ell} = i_\ell, \ldots, X_t = i_0$$

in the data. To ensure that the model defines a high-order Markov chain, the log-likelihood must be maximized with respect to the constraints (9) and either (10) or (11).

Two software packages are currently available to maximize this log-likelihood. The first, called MTD, was developed by Raftery and Tavaré (1994) and can be obtained by sending the message "send mtd from general" to statlib@lib.stat.cmu.edu or directly from http://lib.stat.cmu.edu/general. The main drawback of this software is that it uses an optimization routine from the NAG library which is not freely available. Moreover, in some situations, it does not converge to the global maximum of the likelihood.

The second package is the more recent GMTD software described by Berchtold (2001). It can be obtained from http://www.andreberchtold.com/softwares.html or from http://lib.stat.cmu.edu/general. The algorithm used in GMTD is suboptimal since it does not maximize the likelihood with respect to all the parameters simultaneously. On the other hand, it easily handles all the constraints of the MTD model and it yielded very good results empirically. This algorithm is fully described in Berchtold (2001); here we outline it briefly.

The whole set of parameters of the MTD model can be divided into $(m + 1)$ subsets: the m rows of the transition matrix Q and the vector λ. It is possible to increase the log-likelihood by changing only one of these sets, keeping the other m fixed. As a first approximation, we suppose that the matrix Q is constant and we seek to increase the log-likelihood by reevaluating the vector λ. We also activate the constraints (10).

Since the sum of the lag parameters is equal to one, the idea is to balance an increase in one of these parameters with an equal decrease in another parameter. The problem lies in the choice of the two parameters to modify. It is easy to see that

$$\frac{\partial LL}{\partial \lambda_k} = \sum_{i_\ell,\ldots,i_0=1}^{m} n_{i_\ell,\ldots,i_0} \frac{q_{i_k i_0}}{\sum_{g=1}^{\ell} \lambda_g q_{i_g i_0}}, \quad k = 1, \ldots, \ell,$$

is the partial derivative of the log-likelihood with respect to the kth lag parameter. The partial derivative gives a measure of the local impact produced by the change of one parameter upon the log-likelihood. Since all parameters belong to the continuous space [0, 1], all derivatives are nonnegative. The best solution is then to increase the parameter corresponding to the largest derivative and to decrease the parameter corresponding to the smallest derivative. If we denote by λ_+ the parameter to be increased, denote by λ_- the parameter to be decreased and denote by δ the amount of change, the reestimation of λ is achieved by replacing λ_+ by $(\lambda_+ + \delta)$ and λ_- by $(\lambda_- - \delta)$.

The log-likelihood (28) is then computed with the reestimated lag vector. If the new value is larger than the previous one, the new vector λ is accepted and the procedure stops. If not, δ is divided by 2 and the procedure iterates. As δ becomes smaller than a fixed threshold, the procedure stops, even if the vector λ was not reevaluated. The log-likelihood achieved through this procedure is greater than or equal to the previous value.

The same method is applied to each row of the transition matrix Q. The corresponding partial derivatives are

$$\frac{\partial LL}{\partial q_{i_k i_0}} = \sum_{i_\ell,\ldots,i_0=1}^{m} n_{i_\ell,\ldots,i_0} \frac{\lambda_k}{\sum_{g=1}^{\ell} \lambda_g q_{i_g i_0}}.$$

Maximum likelihood estimation for the whole model is then performed iteratively as follows:

1. Initialization
   - Choose initial values for all parameters.
   - Choose a value for δ and a criterion to stop the algorithm.
2. Iterations
   - Reestimate the vector λ by modifying two of its elements.
   - Reestimate the transition matrix Q by modifying two elements of each row.
3. Stopping criterion
   - If the increase of the log-likelihood since the last iteration is greater than the stop criterion, go back to step 2.
   - Otherwise, end the procedure.

The same method can be used to estimate the MTDg model in which a different matrix represents the transition probabilities between each lag and the present. In this case, ℓ different transition matrices have to be reevaluated during the second step of the algorithm, instead of just one. See Berchtold (2001) for more details.

Like other iterative methods, this algorithm does not ensure convergence toward the global maximum of the log-likelihood. In some cases, it can converge to a local maximum or even a saddle point. To maximize the probability of reaching the global maximum, a good choice of the initial values is very important. Berchtold (2001) proposed computing a measure of the strength of the association between each lagged value and the present one, and using this information to choose starting values for the algorithm.

The choice of initial values for the vector λ of lag parameters has also been addressed by Mehran (1989a) and Berchtold (1998). Mehran (1989a) computed an empirical estimate of the autocorrelations of his data and used the Yule–Walker equations (15) to find a first estimation of λ. Berchtold (1998) considered a slightly different problem. The transition matrix Q is considered fixed and only the lag parameters can vary. In this case, using an association measure like Theil's u and computing λ_k parameters proportional to it yielded good performance.

The same procedure was also used to compute the spatial MTD model described in Section 3.6, but the log-likelihood used in this case requires some explanation. If the summation of (28) is taken over the entire set of variables X, this equation is no longer the log-likelihood of the model, but a pseudo-log-likelihood (Besag 1975, 1977). Another possibility is to take the summation over a subset $\tilde{X}$ of X defined such that no variables in $\tilde{X}$ are neighbors. Unfortunately, this method does not take into account all the information provided by the data, and in most situations it is preferable to maximize the pseudo-log-likelihood of the model.

4.2 Identifiability of a MTD Model

It is of interest to know if, for a particular data set and a given order, there is a unique best MTD model. Haney (1993) showed that if the model is estimated by maximization of the log-likelihood, the resulting estimate is unique when the following two conditions are satisfied:

1. All the elements of the transition matrix Q are strictly positive.
2. The rows of the transition matrix Q are not all identical.

The first condition is a sufficient condition to ensure that the process has a unique limiting distribution. If the second condition is not satisfied, the MTD model reduces to the independence model.

4.3 Minimum χ^2 Estimation

Different alternatives to the maximization of the log-likelihood have been introduced to estimate the MTD model. Raftery and Tavaré (1994) proposed minimizing the quantity

$$K^2 = \sum_{i_\ell,\ldots,i_0=1}^{m} \frac{(n_{i_\ell,\ldots,i_0} - e_{i_\ell,\ldots,i_0})^2}{e_{i_\ell,\ldots,i_0}} \tag{29}$$

with

$$e_{i_\ell,\ldots,i_0} = n_{i_\ell,\ldots,i_1,+}\, p(i_0|i_\ell,\ldots,i_1),$$

where $n_{i_\ell,\ldots,i_0}$ is the number of sequences of the form

$$X_{t-\ell} = i_\ell, \quad \ldots, \quad X_t = i_0$$

in the data,

$$n_{i_\ell,\ldots,i_1,+} = \sum_{i_\ell,\ldots,i_1=1}^{m} n_{i_\ell,\ldots,i_0}, \quad \forall\, i_0,$$

and $p(i_0|i_\ell,\ldots,i_1)$ denotes the conditional transition probability corresponding to $i_\ell,\ldots,i_0$ given by the MTD model. The sum in (29) is taken only over the combinations of $i_\ell,\ldots,i_0$ for which $n_{i_\ell,\ldots,i_0} > 0$.

Billingsley (1961) proved that if the process is correctly described by an ℓth-order Markov chain, then, asymptotically, K^2 has a χ^2 distribution as the

total number of data $n \to \infty$. This χ^2 approach is an interesting alternative to the maximization of the log-likelihood. In practice, Raftery and Tavaré (1994) reported that they used a numerical procedure from the NAG library, which is unfortunately not freely available. The development of a more usable procedure would be of great interest.

4.4 Generalized Linear Interactive Modeling Analysis of the MTD Model

Raftery and Tavaré (1994) have shown that when the number of values taken by the random variable X_t is $m = 2$, the MTD model can be fitted using an iterative procedure in GLIM (generalized linear interactive modeling; see, e.g., Healy, 1988 and Francis et al., 1993). Consider for instance the case $\ell = 2$, and write $\lambda_1 = \lambda$ and $\lambda_2 = 1 - \lambda$. The log-likelihood (28) can then be rewritten as

$$LL = \sum_{i_2,i_1=1}^{2} \left\{ \sum_{i_0=1}^{2} n_{i_2,i_1,i_0} \log\bigl(p(i_0|i_2,i_1)\bigr) \right\},$$

and, for $X_t = 1$, the $p(i_0|i_2,i_1)$ terms are

$$p(1|i_2,i_1) = \begin{cases} q_{11}, & i_2 = 1,\ i_1 = 1, \\ \lambda\, q_{11} + (1-\lambda)\, q_{21}, & i_2 = 2,\ i_1 = 1, \\ (1-\lambda)\, q_{11} + \lambda\, q_{21}, & i_2 = 1,\ i_1 = 2, \\ q_{21}, & i_2 = 2,\ i_1 = 2, \end{cases}$$

where the probabilities q_{ij} are the elements of the transition matrix Q. When either λ or the probabilities q_{ij} are known, the probabilities $p(1|i_2,i_1)$ are linear in the other parameters. If λ is known, we can define two covariates

$$x_1' = (1\ \lambda\ 1-\lambda\ 0),$$
$$x_2' = (0\ 1-\lambda\ \lambda\ 1),$$

and, using q_{11} and q_{21} as coefficients, we obtain the regression

$$p(1|i_2,i_1) = x_1 q_{11} + x_2 q_{21}. \tag{30}$$

If q_{11} and q_{21} are known, we can define the covariates

$$x_3' = (0\ q_{11} - q_{21}\ q_{21} - q_{11}\ 0),$$
$$x_4' = (q_{11}\ q_{21}\ q_{11}\ q_{21}),$$

and we have

$$p(1|i_2,i_1) = x_3\lambda + x_4. \tag{31}$$

Using (30) and (31) iteratively, we can compute an estimate of the parameters (Aitkin, Anderson, Francis and Hinde 1989, Chapter 6). Both are estimated via maximum likelihood for a generalized linear model with binary response and identity link function.

The same method can be generalized to the cases where $m = 2$ and $\ell > 2$. However, it seems that it cannot easily be extended to $m > 2$, and it is therefore less general than the approaches based on the numerical maximization of the likelihood or the minimization of χ^2.

4.5 EM Algorithm

Le, Martin and Raftery (1996) proposed using the expectation-maximization (EM) algorithm to compute the Gaussian mixture transition distribution model given by (23). This is a two-step iterative algorithm (Baum, 1971; Dempster, Laird and Rubin, 1977; McLachlan and Krishnan, 1996). In the EM algorithm, the potentially observable data are assumed to consist of an observed part X and a missing part Z. During the E *step*, the missing data Z are replaced by their expectation conditionally on X and on the parameters of the model. During the M *step*, the parameters are estimated by maximization of the resulting expected complete-data log-likelihood function. Estimates of parameters are then obtained by iterating these two steps until convergence.

Let $X = (x_1, \ldots, x_n)$ denote the observations and let $Z = (Z_1, \ldots, Z_n)$ denote the missing data, where Z_t is a vector of size $(\ell + 2)$ that has a jth component equal to 1 if x_t comes from the jth component of the conditional cumulative distribution function, and 0 otherwise. Since the Z_i are mutually independent, and X and Z are also independent, the log-likelihood for the complete data (X, Z), conditional upon the first ℓ observations, is

$$\begin{aligned} CLL &= \sum_{t=\ell+1}^{n} \sum_{g=0}^{\ell+1} \log\bigl[\lambda_g^{z_{t,g}}\, f_g(x_t|x_\ell^{t-1}, z_0, \ldots, z_{\ell+1})^{z_{t,g}}\bigr] \\ &= \sum_{t=\ell+1}^{n} \Biggl[\sum_{g=0}^{\ell+1} z_{t,g} \log(\lambda_g) - \sum_{g=0}^{\ell+1} z_{t,g} \log(\sigma_g) \\ &\qquad - z_{t,0} \frac{(x_t - \sum_{g=1}^{\ell} \phi_{0g} x_{t-g})^2}{2\sigma_0^2} \\ &\qquad - \sum_{g=1}^{\ell} z_{t,g} \frac{(x_t - \phi_g x_{t-g})^2}{2\sigma_g^2} \\ &\qquad - z_{t,\ell+1} \frac{x_t^2}{2\sigma_{\ell+1}^2} \Biggr] \\ &\qquad - \frac{n-\ell}{2} \log(2\pi). \end{aligned} \tag{32}$$

During the E step of the algorithm, the parameters of the model are supposed known and the missing data Z are replaced by their expectations. During the M step, the missing data Z are supposed known and the parameters are reestimated by maximization of the log-likelihood function (32). See Le, Martin and Raftery (1996) and Berchtold and Raftery (1999) for more details.

Le, Martin and Raftery (1996) performed numerical simulations to test the quality of the estimations obtained through this method. Their results indicate good performance of the estimators in terms of both bias and variability. The more general MAR model given by (25) can be estimated using a similar EM algorithm.

As previously noted, the log-likelihood given by (32) is the log-likelihood of the complete data, both observed (X) and unobserved (Z). When comparing the GMTD model with other methods, it is useful to compute a log-likelihood for the observed data only. This can be done after the EM algorithm has converged using the quantity

$$\begin{aligned} LL = \sum_{t=\ell+1}^{n} \log\Bigg[& \frac{\lambda_0}{\sigma_0} \exp\left\{ -\frac{(x_t - \sum_{g=1}^{\ell} \phi_{0g} x_{t-g})^2}{2\sigma_0^2} \right\} \\ & + \sum_{g=1}^{\ell} \frac{\lambda_g}{\sigma_g} \exp\left\{ -\frac{(x_t - \phi_g x_{t-g})^2}{2\sigma_g^2} \right\} \\ & + \frac{\lambda_{\ell+1}}{\sigma_{\ell+1}} \exp\left\{ -\frac{x_t^2}{2\sigma_{\ell+1}^2} \right\} \Bigg] \\ & - \frac{n-\ell}{2} \log(2\pi). \end{aligned} \tag{33}$$

This EM approach could also be used to estimate the MTD model for a discrete state space, as well as more general MTD models.

Although we have not investigated it, this general formulation could also provide a way to carry out fully Bayesian estimation using Markov chain Monte Carlo (Gilks, Richardson and Spiegelhalter,1996). A Metropolis–Hastings algorithm could be used, updating one or several parameters and Z values at a time. The full conditional distributions will be known for all the Z and possibly most or all of the parameters, depending on the model being estimated, and Gibbs sampling steps could be used for these missing data and parameters.

Note that the solution space for the class of MTD models can be very large and highly nonlinear, especially in the continuous case. Even if the EM algorithm works well, it can converge to a local optimum rather than to the global one. Thus the choice of the set of initial parameter values can be crucial. One method is to explore the solution space using a genetic algorithm [Holland (1975)], the function to optimize being the log-likelihood (33). However, genetic algorithms tend rapidly to find the region containing the solution, but then are slow to converge. So one possibility is to estimate the model through a two-stage procedure:

1. Use a genetic algorithm to determine an initial solution.
2. Improve this solution by using the EM algorithm.

This method combines advantages of both algorithms, with a good probability of finding the global maximum of the log-likelihood in a limited number of iterations. The results of Section 5.5 were obtained using this method, which worked well in that case. The EM algorithm itself can be slow in converging close to the solution, while Newton–Raphson methods tend to be fast. Thus the algorithm might be further accelerated by using Newton–Raphson close to the solution. However, excessive computer time has not been a problem in the examples we have worked on, so this might not be worth the additional trouble.

5. APPLICATIONS

The class of MTD models has been used in a range of applications including the analysis of wind speed and direction, DNA sequences, social behavior and financial series. In this section, we present synthetically the different types of applications and we summarize the most interesting results. The different models are compared using the BIC criterion defined by (6).

5.1 Wind Modeling

We know of four applications of the MTD model to wind data. Raftery (1985a) analyzed time series of wind speeds, while Craig (1989), Raftery and Tavaré (1994) and MacDonald and Zucchini (1997) analyzed wind directions.

Raftery (1985a) considered a time series of 672 hourly wind speeds at Belmullet, Ireland, and wanted to assess the amount of electricity that could be generated from wind power. The data are coded into four states ranging from no power to excessively high winds. The fully parameterized Markov chains of orders 0–4 are compared to the corresponding MTD models. The best result in terms of BIC is achieved by the third-order MTD model, but it is interesting to note that in this case both the second- and fourth-order MTD

models also had better BIC values than the best fully parameterized Markov chain. The lag parameters of the best model decrease with lag: $\lambda_1 = 0.629$, $\lambda_2 = 0.206$ and $\lambda_3 = 0.165$. The transition matrix Q is

$$Q = \begin{bmatrix} 0.837 & 0.163 & 0 & 0 \\ 0.058 & 0.854 & 0.088 & 0 \\ 0 & 0.113 & 0.847 & 0.040 \\ 0 & 0 & 0.116 & 0.884 \end{bmatrix}.$$

Interestingly, even though the Q matrix gives large weight to the probabilities of the diagonal, the model of Pegram (Section 6.1) does not fit as well as the MTD model for these data.

Raftery and Tavaré (1994) analyzed another set of wind data, a sequence of length 4344 that gives the hourly wind directions at Roche's Point, Ireland, for November 1961–April 1962. The data are coded in five categories, the first denoting the absence of wind and the four others representing different directions.

Even though the data set is moderately large, it is not large enough to fit high-order fully parameterized Markov chains, and among these models, the first-order one is preferred. On the other hand, the use of the MTD model permits the analysis of long dependency patterns. Models were fitted for orders 1–10, and the seventh-order MTD model yielded the best result.

The lag parameters of this model do not decrease strictly with lag; the smallest λ was $\lambda_4 = 0.018$. The authors then decided to impose the constraint $\lambda_4 = \lambda_5 = \lambda_6 = 0$ and to recompute the seventh-order model. The new model proved to be better than all others considered. It has lag parameters $\lambda_1 = 0.598$, $\lambda_2 = 0.245$, $\lambda_3 = 0.1$ and $\lambda_7 = 0.057$. [Raftery and Tavaré (1994) gave $\lambda_3 = 0.001$, which is incorrect.]

As already noted with the epileptic data given in Section 1, this example shows that the MTD model is particularly well suited for the analysis of dependences at higher lags. Here, even though the first three lags account for 94.3% of the variability in predicting the present value, adding the seventh lag improves the results. This would not have been possible with the corresponding fully parameterized Markov chain.

MacDonald and Zucchini (1997) presented a sequence of hourly wind directions at Koeberg, South Africa, covering the period May 1, 1985–April 30, 1989, for a total length of 35,064 (The Koeberg data used in Section 1 are taken from this time series). The data are coded into 16 directions. In contrast with the data of Raftery and Tavaré (1994), there is no "no wind" state. Again, in this case, the MTD model (of second-order) is preferred to the first-order Markov chain, but the gain is not as large as in the preceding example. The weight of the first lag is $\lambda_1 = 0.9125$ and the transition matrix is close to the first-order Markov chain, suggesting that adding more lags would not greatly improve the model.

5.2 Social Behavior

The transition process between different social behaviors can also be well represented by Markovian models. The MTD model has been used at least four times in this context. Raftery (1985a) reanalyzed a set of data previously published in Katz and Proctor (1959) and Bishop, Fienberg and Holland (1975). This is a sample of two-step transitions giving the relationships between 300 students taken at intervals of 2 months. There are three states: mutual, one-way and indifferent. The first-order model is rejected, and among the second-order models, the MTD model performs considerably better than the full Markov chain according to BIC. This result is due in part to the parsimony of the MTD model which in this case has only 5 independent parameters, compared with 12 for the second-order Markov chain. The MTD lag parameters are $\lambda_1 = 0.754$ and $\lambda_2 = 0.246$, and the transition matrix is

$$Q = \begin{bmatrix} 0.581 & 0 & 0.419 \\ 0.133 & 0.545 & 0.322 \\ 0 & 0.093 & 0.907 \end{bmatrix}.$$

Note that, according to the convention of Bishop, Fienberg and Holland (1975), the 0's are not counted as parameters.

Raftery (1985a) also reanalyzed the data of Logan (1981). This is a sample of two-step transitions between the main occupations of 9170 U.S. physicists taken at intervals of 2 years. The states are management, research and teaching. Again, the second-order MTD model proved to provide the best possible Markovian modeling.

Mehran (1989a, b) used the MTD model to analyze the probability of unemployment in the United States. His data were collected using a rotating sample scheme. Each person in the survey was observed during 4 consecutive months, then dropped from the sample for 8 months, and finally observed again during 4 months. Three states were considered: employed, unemployed and out of the labour force. More complete results appear in Mehran (1989a) where a three-lag model was fitted to the first four periods of the rotating sampling scheme (December 1982–March 1983). Here we analyze only the data on people who were in

the labor force for the entire period, so that we have only two states to consider. The model has lag parameters $\lambda_1 = 0.730$, $\lambda_2 = 0.161$ and $\lambda_3 = 0.109$, and transition matrix

$$Q = \begin{pmatrix} 0.989 & 0.011 \\ 0.233 & 0.767 \end{pmatrix}.$$

The resulting transition probabilities are

X_{t-3}	X_{t-2}	X_{t-1}	X_t = E	X_t = U
E	E	E	0.989	0.011
U	E	E	0.907	0.093
E	U	E	0.868	0.132
U	U	E	0.785	0.215
E	E	U	0.437	0.563
U	E	U	0.355	0.645
E	U	U	0.317	0.683
U	U	U	0.233	0.767

where E denotes employment and U denotes unemployment. The model fits the data remarkably well. The probability of being employed at time t depends strongly on the probability of being employed at time $t-1$. When $X_{t-1} = \mathrm{E}$, the probability that $X_t = \mathrm{E}$ is at least 0.785 whatever X_{t-2} and X_{t-3} are. On the other hand, this probability is always lower than 0.5 when $X_{t-1} = \mathrm{U}$. It is also interesting to note that the model explains the present as a function of the elapsed time since the last employment period. For instance, the probability of being employed at time t given $X_{t-1} = \mathrm{U}$, $X_{t-2} = \mathrm{E}$ and $X_{t-3} = \mathrm{U}$ (0.355) is greater than the probability of being employed when $X_{t-1} = \mathrm{U}$, $X_{t-2} = \mathrm{U}$ and $X_{t-3} = \mathrm{E}$ (0.317).

A MTD-like model has been used for the analysis of the mobility of insured individuals between different types of health insurance plans in Switzerland (Berchtold, 1997, 1998). The cost of each plan was also investigated. The data consist of a set of 4300 Swiss people observed during the period 1991–1994. They are described by a set of 13 variables including sex, age, type of health insurance, number of health bills each year, yearly total health cost and level of education. Transition matrices were computed between each variable and the type of insurance. Then an association measure was computed for each matrix, and the ones having the greatest predictive power for the type of insurance were combined through a MTD model to obtain matrices explaining the choice of the type of insurance with two variables simultaneously. Results obtained through this method proved to be consistent with theoretical predictions. The same method was applied to establish a link between the same set of explanatory variables and the yearly total health cost.

5.3 DNA Sequences

Raftery and Tavaré (1994) used Markov chains to fit two sequences of mouse DNA. Three decompositions of nucleotides were considered: the complete set of four bases {A, C, G, T}, the {A/G, C, T} decomposition and the binary purine–pyrimidine alphabet where each base is recoded as either purine ({A, G}) or pyrimidine ({C, T}). For both sequences, the complete four-base alphabet leads to a first-order Markov chain that cannot be improved upon by a MTD model. In the case of the three-state alphabet {A/G, C, T}, the second-order MTD model is chosen. Finally, for the purine–pyridine case, the first sequence is best modeled by a fully parameterized second-order Markov chain, the second-order MTD being the second best. The second sequence leads to the second-order MTD.

One question raised by this analysis is the fact that a time-series model is used to represent sequences of data in which there is no explicit order. Berchtold (2001) reanalyzed one of these two sequences in its complete four letter alphabet form. A first analysis shows that the sequence can be analyzed equally well by considering an ordering from the left to the right or from the right to the left, suggesting that a time-series model may not be the best solution in this case. Then the spatial MTD model discussed in Section 3.6 was applied, using either one or two bases on each side of the focal base. The best model is obtained using one base on each side and a different transition matrix to represent transitions from each side of the focal base. Denoting by L the base on the left and by R the base on the right, this model has parameters

$$\lambda_{\mathrm{L}} = 0.4598, \qquad \lambda_{\mathrm{R}} = 0.5402,$$

$$Q_{\mathrm{L}} = \begin{bmatrix} 0.2039 & 0.1311 & 0.5883 & 0.0767 \\ 0.3654 & 0.2674 & 0 & 0.3672 \\ 0.2263 & 0.0981 & 0.5565 & 0.1191 \\ 0.1470 & 0.1324 & 0.5222 & 0.1984 \end{bmatrix},$$

$$Q_{\mathrm{R}} = \begin{bmatrix} 0.2136 & 0.5124 & 0.1037 & 0.1703 \\ 0.1878 & 0.5059 & 0.0461 & 0.2602 \\ 0.4086 & 0 & 0.2323 & 0.3591 \\ 0.1386 & 0.5462 & 0.0896 & 0.2256 \end{bmatrix}.$$

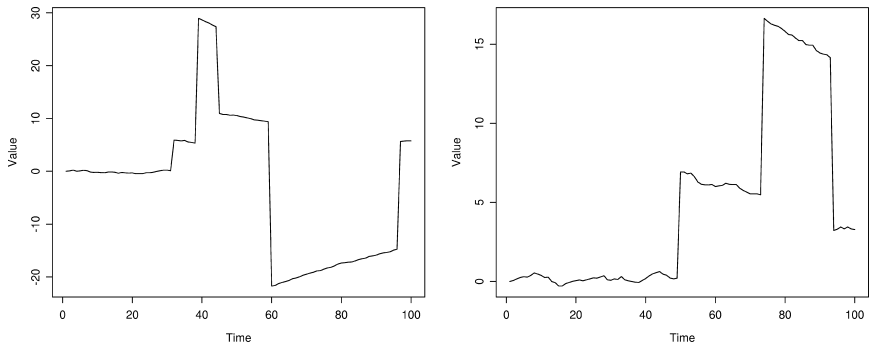

FIG. 4. *Two trajectories of an MTD model. Note the clear change-point-like behavior.* Source: *Raftery* (1993).

5.4 Change Points, Bursts, Outliers and Flat Stretches

Raftery (1993) and Le, Martin and Raftery (1996) showed that the class of MTD models is able to represent change points in time-series without explicitly including them in the model; it can also represent flat stretches, bursts of high volatility or variance, and outliers. A big advantage of this is that it is not necessary to define a model specifically for each type of data set analyzed or for each type of nonlinear or non-Gaussian behavior expected. Figure 4 shows trajectories of a MTD model which can reproduce change-point-like behavior.

Le, Martin and Raftery (1996) provided two examples of the use of the Gaussian MTD model (see Section 3.5). The first is the daily closing price of IBM common stock from May 17, 1961 to November 2, 1962. The second example is a series of consecutive hourly viscosity readings from a chemical process. In both cases, the GMTD outperformed the classical ARIMA model. Figure 5 presents two prediction intervals for the viscosity series and compares the AR and GMTD models. The 90% intervals are similar for the MTD and the AR models, but the intervals obtained using the MTD model are narrower in the 60% case. Overall, the GMTD model is preferred.

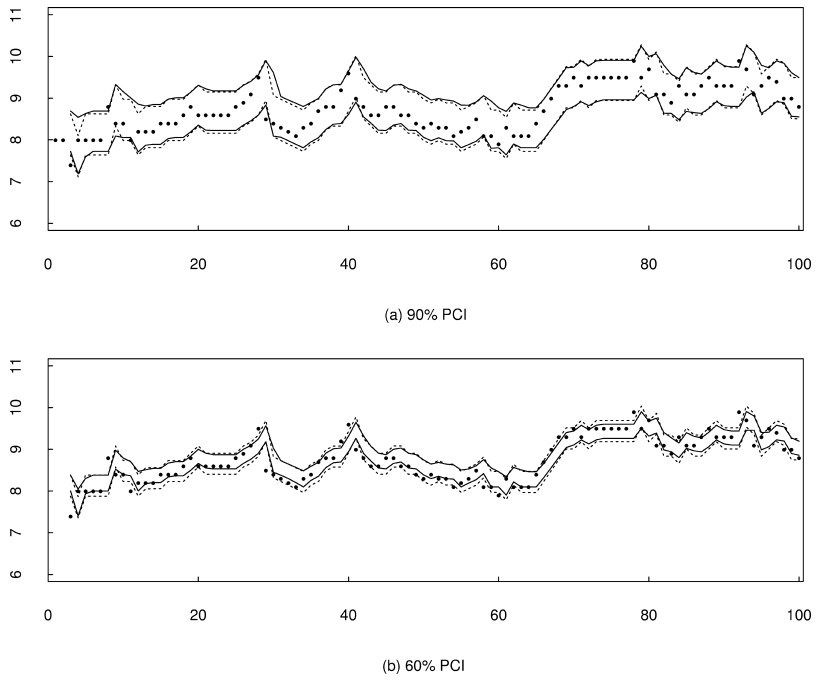

FIG. 5. (a) 90% *and* (b) 60% *predictive intervals for a chemical viscosity series. The dots denote the original observations, the dashed lines are for the AR-based predictive intervals and the solid lines are for the Gaussian MTD-based predictive intervals.* Source: *Le, Martin and Raftery* (1996).

Simulations were also performed to compare the ability of the MTD and AR model to capture sudden bursts of activity in a time series. Once again, the MTD model was preferred. This kind of behavior has often been modeled using bilinear time-series models (e.g., Subba Rao and Gabr, 1984), and these results suggest that MTD models may be able to carry out some of the functions of bilinear time-series models without needing to specify a bilinear structure.

5.5 Financial and Economic Time Series

Financial time series have proven to be difficult to model. They often exhibit a near-random behavior characterized by nonstationarity and heteroskedasticity, and require the use of specially designed econometrics models (Bollerslev, Chou and Kroner, 1992; Hamilton 1994). For example, Figures 6 and 7 show the evolution of the closing price of Eastman Kodak shares from May 12, 1998 to May 2, 2000 and the corresponding first-differenced series. This series exhibits a negative trend with short periods of very high variance (volatility).

The usual candidates to represent this kind of behavior are the autoregressive conditional heteroskedasticity model (ARCH; Engle, 1982) and the generalized autoregressive conditional heteroskedasticity model (GARCH; Bollerslev, 1986; Bollerslev, Chou and Kroner, 1992). The general principle is to model not only the level of the random variable, but also its variance. Let X_t be a continuous random variable and consider

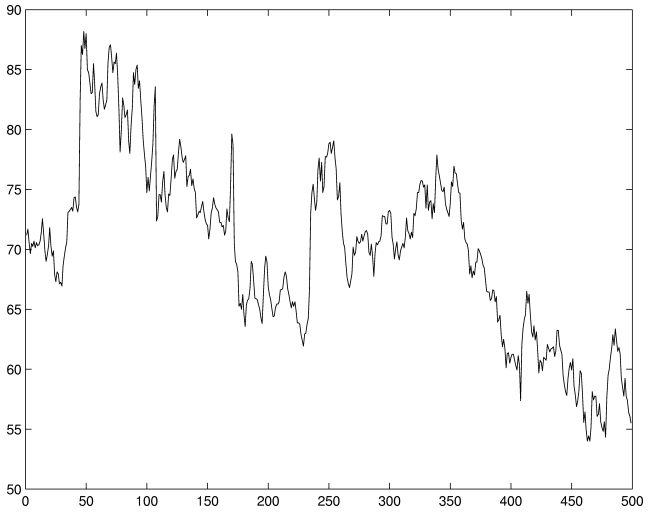

FIG. 6. *Closing price of Eastman Kodak shares from May 12, 1998 to May 2, 2000* (499 *observations*).

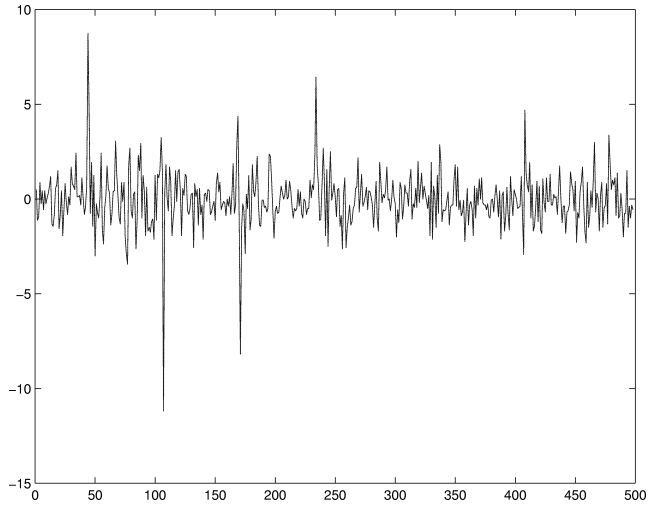

FIG. 7. *Eastman Kodak first-differenced series.*

the model

$$X_t = c + \sum_{j=1}^{\ell} \lambda_j X_{t-j} + \varepsilon_t,$$

where ε_t is white noise with expectation 0 and variance σ^2. To take heteroskedasticity into account, the square of ε_t is itself represented as an autoregressive process of order q,

$$\varepsilon_t^2 = d + \sum_{j=1}^{q} \varphi_j \varepsilon_{t-j}^2 + u_t, \tag{34}$$

where u_t is white noise. Equation (34) defines an ARCH(q) process. This process is often of high order and it is useful to consider an infinite-lag representation. This leads to the equation for the variance

$$\sigma_t^2 = \omega + \sum_{j=1}^{q} \varphi_j \varepsilon_{t-j}^2 + \sum_{j=1}^{p} \zeta_j \sigma_{t-j}^2. \tag{35}$$

Equation (35) defines a GARCH(p, q) model. There have been many generalizations of this class of models, including the exponential GARCH model [EGARCH(p, q)], which uses not only the magnitude, but also the sign of ε to predict the variance, and the ARCH-in-mean model in which an increase in the conditional variance is related to a change in the conditional mean.

The ARCH-type models can be very good at the representation of series like the one in Figure 6, but, as noted by Wong and Li (2000), the GMTD family also has the ability to represent such behavior. Table 4 compares different models for the Eastman Kodak series. All the models were estimated using the

TABLE 4
Comparison of different models for the Eastman Kodak series

Model for the mean	Model for the variance	LL	Number of parameters	BIC
AR(1)	—	−876.34	2	1765.09
AR(1)	GARCH(1, 1)	−844.86	5	1720.74
AR(1)	GARCH(2, 2)	−841.45	7	1726.33
AR(1)	EGARCH(2, 2)	−829.85	9	1715.54
AR(2)	—	−872.34	3	1763.29
AR(2)	GARCH(1, 1)	−843.95	6	1725.13
AR(2)	GARCH(2, 2)	−840.25	8	1730.14
AR(2)	EGARCH(2, 2)	−827.74	10	1717.53
RWGMTD(2)	—	−810.60	6	**1658.43**
MAR(3, 2, 2, 2)	—	−807.28	14	1701.42

NOTE: The log-likelihood of each model has 495 components. AR(ℓ) stands for an ℓth-order autoregressive model, RWGMTD(ℓ) is an ℓth-order random walk GMTD and MAR(k, $p_1, \ldots, p_k$) is a k-component MAR model with orders $p_1, \ldots, p_k$.

last 495 data points, conditionally upon the first four observations.

Here, the best model is the second-order random walk GMTD

$$\begin{aligned} F(x_t|x_0^{t-1}) &= 0.9651\Phi\left(\frac{x_t - 1.1274x_{t-1} + 0.1274x_{t-2}}{1.1747}\right) \\ &\quad + 0.0218\Phi\left(\frac{x_t - x_{t-1}}{0.4146}\right) \\ &\quad + 0.0131\Phi\left(\frac{x_t - x_{t-2}}{9.4837}\right). \end{aligned}$$

This model achieves a larger log-likelihood than the best GARCH model does, and since it does not have too many parameters, it is the best overall. The use of a MAR model improves the log-likelihood slightly, but at the cost of using a much greater number of parameters.

Another example is provided by the U.S. annual consumer price inflation level from 1821 to 1999. Figures 8 and 9 show the original series and the first-differenced series. Table 5 summarizes the results from different models for this series.

Once again, a second-order random walk GMTD yields the best result, this time with an independent

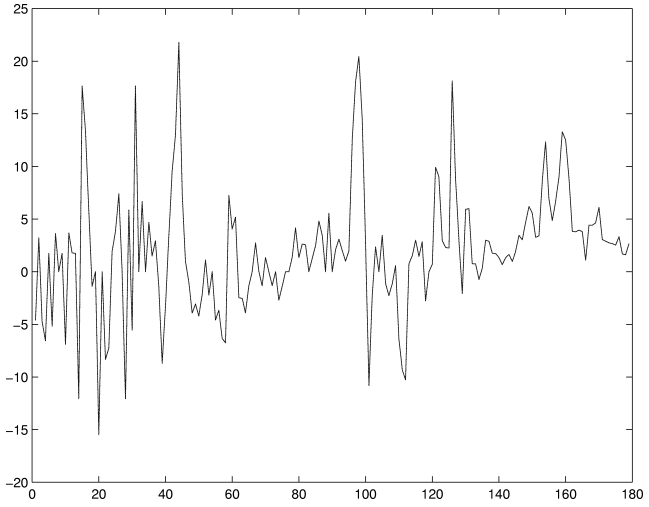

FIG. 8. *U.S. annual consumer price inflation level from 1821 to 1999* (179 *observations*).

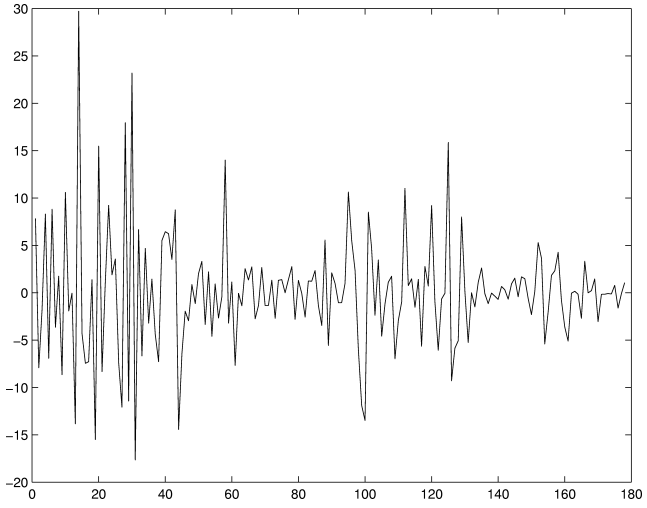

FIG. 9. *First-differenced series of the U.S. annual consumer price inflation level series.*

TABLE 5
Comparison of different models for the U.S. inflation series

Model for the mean	Model for the variance	LL	Number of parameters	BIC
AR(1)	—	−536.46	2	1083.24
AR(1)	GARCH(1, 1)	−508.32	5	1042.46
AR(1)	GARCH(1, 2)	−500.47	6	1031.93
AR(1)	GARCH(2, 2)	−500.39	7	1036.93
RWGMTDi(2)	—	−485.39	8	**1012.10**
MAR(4, 1, 1, 1, 1)	—	−497.28	15	1072.02
MAR(4, 2, 2, 2, 2)	—	−473.64	17	1035.08

NOTE: The log-likelihood of each model has 175 components. AR(ℓ) stands for an ℓth-order autoregressive model, RWGMTDi(ℓ) is an ℓth-order random walk GMTD with an independent term and MAR(k, $p_1, \ldots, p_k$) is a k-component MAR model with orders $p_1, \ldots, p_k$.

term. It is written

$$F(x_t|x_0^{t-1}) = 0.2424\Phi\left(\frac{x_t - 1.6016x_{t-1} + 0.6016x_{t-2}}{5.0806}\right) + 0.498\Phi\left(\frac{x_t - x_{t-1}}{2.5002}\right) + 0.2323\Phi\left(\frac{x_t - x_{t-2}}{10.1}\right) + 0.0273\Phi\left(\frac{x_t}{0.0122}\right).$$

As in the previous example, the GARCH models are outperformed by the MTD specifications. Moreover, the best overall model is not the one having the best log-likelihood (a four-component MAR model of order 2), because the number of parameters is too large.

These examples demonstrate that the GMTD model can handle series with heteroskedasticity well without having to model this particular behavior explicitly. This exhibits the flexibility and power of MTD modeling once again, and thus defines an alternative to the widely used GARCH models that may be of some interest.

5.6 Biological Applications

The MTD model has been used in the literature for the analysis of a range of other data sets. Raftery and Tavaré (1994) and Berchtold (2001) analyzed the song of the wood pewee. The sequences generated by this New England song bird are characterized by often repeated patterns or phrases. In its generic form as discussed here, the MTD is characterized by a lack of interaction between past values in their effect on the present state. In the wood pewee song, however, such interactions are not only present, but are a major feature of the data. A generalization of the MTD model that explicitly modeled the bird's main phrases performed very well. This idea might be relevant for other data where strong patterns are a feature, such as coding regions in DNA or protein sequences.

Berchtold (1998) also used a MTD model to study the spatial distribution of *Carex arenaria*, a European marsh plant. More generally, he proposed using the MTD principle to create transition matrices between several explanatory and/or explained variables, starting with single matrices between only one explanatory and one explained variable. The statistical characteristics of this general model are not well known yet.

6. DISCUSSION

In this section, we discuss other solutions proposed in the literature for the modeling of high-order Markov chains or, more generally, for the modeling of non-Gaussian time series. We also discuss further applications of the MTD model in hidden Markov models and the use of covariates.

6.1 Other Models for High-Order Markov Chains

We know of only three other classes of models for high-order Markov chains. The first is due to Jacobs and Lewis (1978c) and Pegram (1980) and generalizes a model previously independently proposed by several authors including Lloyd (1977) and Jacobs and Lewis (1978a, b). This model is a special case of the MTD model, where the transition matrix Q takes the form

$$Q = \theta I + (1-\theta)\iota\pi'$$
$$= \begin{pmatrix} \theta+(1-\theta)\pi_1 & (1-\theta)\pi_2 & \cdots & (1-\theta)\pi_m \\ (1-\theta)\pi_1 & \theta+(1-\theta)\pi_2 & \cdots & (1-\theta)\pi_m \\ \vdots & & \ddots & \vdots \\ (1-\theta)\pi_1 & \cdots & \cdots & \theta+(1-\theta)\pi_m \end{pmatrix},$$

where θ is a parameter, I is the $m \times m$ identity matrix, ι is an m vector of 1's and $\pi' = (\pi_1, \ldots, \pi_m)$ is the limiting distribution of the first-order Markov chain. This model has only $(m + \ell - 1)$ independent parameters, which is generally more parsimonious than the corresponding MTD model. On the other hand, the constraints on the transition matrix Q are very restrictive. The probability of being in a particular state k at time t depends only on the probability of being in the same state in the past. For every other value of the lag, the probability is identical, that is, $(1-\theta)\pi_k$. Raftery (1985a, b) proved that this model can represent only a subset of the autocorrelation range of the MTD model, and so it is useful only in specific situations. Further non-Markovian generalizations were also proposed by Jacobs and Lewis (1983).

The second class of high-order Markov chain models is due to Logan (1981). Two models were presented: one constrained and the other unconstrained. These models imply a larger number of parameters than the MTD model and we do not know of any further developments or applications.

The third class of models for high-order Markov chains is the variable length Markov chain model (VLMC) (Weinberger, Rissanen and Feder, 1995; Bühlmann and Wyner, 1999). The principle of this class of models is to explore all branches of an ℓth order Markov chain and to aggregate branches that present similar probability distributions. For instance, consider a random variable X taking values in $\{1, 2\}$ and the following third-order transition matrix given in reduced form:

$$R = \begin{array}{ccc|cc} & & & \multicolumn{2}{c}{X_t} \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & 0 & 1 \\ 2 & 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0.25 & 0.75 \\ 2 & 2 & 1 & 0.6 & 0.4 \\ 1 & 1 & 2 & 0.74 & 0.26 \\ 2 & 1 & 2 & 0.73 & 0.27 \\ 1 & 2 & 2 & 0.74 & 0.26 \\ 2 & 2 & 2 & 0.71 & 0.29 \end{array}.$$

The transition matrix R is fully specified by eight independent parameters, but the first and second rows are identical, that is, the probability of X_t is independent of X_{t-3} given $X_{t-1} = X_{t-2} = 1$. Moreover, the last four rows are similar. This suggests reparameterizing R as

$$R' = \begin{array}{ccc|cc} & & & \multicolumn{2}{c}{X_t} \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & q_1 & 1-q_1 \\ 2 & 1 & 1 & q_1 & 1-q_1 \\ 1 & 2 & 1 & q_2 & 1-q_2 \\ 2 & 2 & 1 & q_3 & 1-q_3 \\ 1 & 1 & 2 & q_4 & 1-q_4 \\ 2 & 1 & 2 & q_4 & 1-q_4 \\ 1 & 2 & 2 & q_4 & 1-q_4 \\ 2 & 2 & 2 & q_4 & 1-q_4 \end{array}$$

with $q_1 = 0$, $q_2 = 0.25$, $q_3 = 0.6$ and $q_4 = 0.73$. The transition matrix R' then defines a VLMC model that has only four independent parameters, compared to the eight parameters of the fully parameterized third-order Markov chain. A procedure for determining the best VLMC model is described in Bühlmann and Wyner (1999).

The VLMC model is interesting in that it presents an approach complementary to the MTD. The MTD model implicitly supposes that the process is of full ℓth order and that all branches have to be estimated, albeit parsimoniously. On the other hand, the VLMC model supposes that only a part of the structure of the data is of full ℓth order.

To show this complementarity, we performed the following simulation experiment. First, we considered a random variable taking values in $\{1, 2\}$ and a second-order transition matrix T_1:

$$T_1 = \begin{array}{cc|cc} & & \multicolumn{2}{c}{X_t} \\ X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 0.9 & 0.1 \\ 2 & 1 & 0.6 & 0.4 \\ 1 & 2 & 0.5 & 0.5 \\ 2 & 2 & 0.5 & 0.5 \end{array}.$$

Since the last two rows are identical, this matrix can be described by three independent parameters only. It is a good candidate to be modeled by a VLMC model, namely

$$T_1' = \begin{array}{cc|cc} & & \multicolumn{2}{c}{X_t} \\ X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & q_1 & 1-q_1 \\ 2 & 1 & q_2 & 1-q_2 \\ 1 & 2 & q_3 & 1-q_3 \\ 2 & 2 & q_3 & 1-q_3 \end{array}.$$

TABLE 6
Comparison of the performance of the VLMC and MTD models, which are computed on sequences generated by the transition matrices T_1 and T_2; results for the independence model and for the first- and second-order full Markov chains (MC) are also included

Transition matrix	Number of data	Models				
		Independence	MC 1	MC 2	MTD 2	VLMC
T_1	50	288	353	161	51	158
	100	101	400	39	88	372
	300	1	79	11	109	**800**
	1000	0	0	5	24	**971**
T_2	50	251	296	155	239	73
	100	136	209	40	**527**	89
	300	3	17	42	**903**	35
	1000	0	0	86	**914**	0

NOTE: For each model, the table gives the number of times the model was chosen as the best in 1000 replications according to the BIC criterion. Models that were chosen as best more than 50% of the time are shown in bold. Note that for small sample sizes, the sum of a row is sometimes greater than 1000. This indicates that several models obtained the same best BIC value.

We used T_1 to generate sequences of size 52, 102, 302 and 1002, and we computed five models: the independence model, Markov chains of order 1 and 2, the MTD model of order 2 and the VLMC model using the parameterization of T_1'. This procedure was replicated 1000 times for each sequence length and models were compared on the basis of their BIC values. (To ensure a fair comparison, we conditioned on the first two data points of each sequence so as to have the same number of components in the log-likelihood for each model, namely 50, 100, 300 and 1000.) Then we carried out the same experiment again, this time using the transition matrix T_2, namely

$$T_2 = \begin{array}{cc} & X_t \\ X_{t-2} \quad X_{t-1} & \quad 1 \quad\quad 2 \\ \begin{array}{c} 1 \\ 2 \\ 1 \\ 2 \end{array} \quad \begin{array}{c} 1 \\ 1 \\ 2 \\ 2 \end{array} & \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \\ 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix} \end{array}.$$

Intuitively, it would seem that the VLMC model defined by T_1' should be good when used on sequences generated by T_1, but should not perform well on sequences generated by T_2. On the other hand, the MTD model should obtain good results with T_2. So,we expected to obtain better results by the VLMC on sequences generated by T_1 and by the MTD on sequences generated by T_2. Table 6 summarizes our results in terms of BIC. Note that the MTD and VLMC models have the same number of independent parameters (three), so the BIC comparison between these two models amounts simply to preferring the one with the larger log-likelihood.

In the case of small sequence sizes (50 data points, and also 100 data points in the case of T_1), Table 6 shows a great variability between models. Often, the simple first-order model is chosen. On the other hand, as the sequence length increases, the VLMC model becomes the best choice for sequences generated by T_1 and the MTD model becomes the best choice for sequences generated by T_2. These results corroborate our hypothesis. The VLMC and MTD models are not competitive, but represent complementary solutions for the modeling of high-order dependencies.

This complementarity between MTD and VLMC can also be used to improve the modeling of more complex situations. Consider the following third-order

transition matrix V:

$$V = \begin{array}{ccc|cc} & & & \multicolumn{2}{c}{X_t} \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & q_1 & 1-q_1 \\ 2 & 1 & 1 & q_2 & 1-q_2 \\ 1 & 2 & 1 & q_3 & 1-q_3 \\ 2 & 2 & 1 & q_4 & 1-q_4 \\ 1 & 1 & 2 & q_5 & 1-q_5 \\ 2 & 1 & 2 & q_6 & 1-q_6 \\ 1 & 2 & 2 & q_6 & 1-q_6 \\ 2 & 2 & 2 & q_6 & 1-q_6 \end{array}.$$

The full third-order matrix would have eight independent parameters, but since the last three rows are identical, a VLMC seems a better choice with only six independent parameters. Another solution is to use a MTD model. It cannot take into account the equality of the last three rows, but on the other hand, it has a smaller number of independent parameters (four). An intermediate solution could be to use both a VLMC and a MTD specification to obtain a better mix between the log-likelihood and the number of parameters. We propose to use the following two-step procedure:

1. Search for the best VLMC model for the data and estimate rows which can be simplified.
2. Compute the MTD model and use it to estimate rows which cannot be simplified by the VLMC model.

Using this procedure, the respective qualities of both models are used together. To test this idea, we performed the following simulation experiment. We considered the following four transition matrices:

$$V_1 = \begin{array}{ccc|cc} & & & \multicolumn{2}{c}{X_t} \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & 0.9 & 0.1 \\ 2 & 1 & 1 & 0.6 & 0.4 \\ 1 & 2 & 1 & 0.7 & 0.3 \\ 2 & 2 & 1 & 0.3 & 0.7 \\ 1 & 1 & 2 & 0.8 & 0.2 \\ 2 & 1 & 2 & 0.5 & 0.5 \\ 1 & 2 & 2 & 0.4 & 0.6 \\ 2 & 2 & 2 & 0.1 & 0.9 \end{array},$$

$$V_2 = \begin{array}{ccc|cc} & & & \multicolumn{2}{c}{X_t} \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & 0.9 & 0.1 \\ 2 & 1 & 1 & 0.6 & 0.4 \\ 1 & 2 & 1 & 0.7 & 0.3 \\ 2 & 2 & 1 & 0.3 & 0.7 \\ 1 & 1 & 2 & 0.8 & 0.2 \\ 2 & 1 & 2 & 0.2 & 0.8 \\ 1 & 2 & 2 & 0.2 & 0.8 \\ 2 & 2 & 2 & 0.2 & 0.8 \end{array},$$

$$V_3 = \begin{array}{ccc|cc} & & & \multicolumn{2}{c}{X_t} \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & 0.9 & 0.1 \\ 2 & 1 & 1 & 0.6 & 0.4 \\ 1 & 2 & 1 & 0.7 & 0.3 \\ 2 & 2 & 1 & 0.3 & 0.7 \\ 1 & 1 & 2 & 0.8 & 0.2 \\ 2 & 1 & 2 & 0.5 & 0.5 \\ 1 & 2 & 2 & 0.5 & 0.5 \\ 2 & 2 & 2 & 0.5 & 0.5 \end{array},$$

$$V_4 = \begin{array}{ccc|cc} & & & \multicolumn{2}{c}{X_t} \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & 0.9 & 0.1 \\ 2 & 1 & 1 & 0.6 & 0.4 \\ 1 & 2 & 1 & 0.7 & 0.3 \\ 2 & 2 & 1 & 0.3 & 0.7 \\ 1 & 1 & 2 & 0.8 & 0.2 \\ 2 & 1 & 2 & 0.4 & 0.6 \\ 1 & 2 & 2 & 0.4 & 0.6 \\ 2 & 2 & 2 & 0.4 & 0.6 \end{array}.$$

Note that these four matrices differ only by their last three rows. The V_1 matrix is of full third order, while the V_2, V_3 and V_4 matrices have the same structure as V with the last three rows equal. Each matrix was used to generate 1000 sequences of length 1003. We computed four models on each sequence: the full third-order transition matrix, the third-order MTD, the VLMC having the structure of matrix V and a model mixing MTD and VLMC in which the first five rows correspond to the MTD and the last three rows correspond to the VLMC. Table 7 summarizes the results.

One would not expect the VLMC or the MTD-VLMC models to provide good representations of

TABLE 7
Comparison of the performance of the MTD, VLMC and MTD-VLMC models, which are computed on sequences generated by the transition matrices V_1 to V_4; results for the third-order full Markov chains (MC) are also included

Transition matrix	Models			
	MC 3	MTD 3	VLMC	MTD-VLMC
V_1	15	**985**	0	0
V_2	1	5	**819**	175
V_3	0	330	217	453
V_4	0	133	218	**649**

NOTE: For each model, the table gives the number of times the model was chosen as the best in 1000 replications according to the BIC criterion. The number is shown in bold if the model was best more than 50% of the time.

sequences generated by V_1, since this matrix does not have a VLMC structure. In this case, the MTD model is almost always preferred. Sequences generated using V_2 are generally best modeled by a VLMC model. Finally, the MTD-VLMC model achieves good performances on the last two matrices. This can be explained by looking at the reduced form of the third-order MTD:

$$\text{MTD 3} = \begin{array}{ccc|cc} & & & X_t & \\ X_{t-3} & X_{t-2} & X_{t-1} & 1 & 2 \\ \hline 1 & 1 & 1 & q_{11} & q_{12} \\ 2 & 1 & 1 & \lambda_1 q_{21} + \lambda_2 q_{11} + \lambda_3 q_{11} & \lambda_1 q_{22} + \lambda_2 q_{12} + \lambda_3 q_{12} \\ 1 & 2 & 1 & \lambda_1 q_{11} + \lambda_2 q_{21} + \lambda_3 q_{11} & \lambda_1 q_{12} + \lambda_2 q_{22} + \lambda_3 q_{12} \\ 2 & 2 & 1 & \lambda_1 q_{21} + \lambda_2 q_{21} + \lambda_3 q_{11} & \lambda_1 q_{22} + \lambda_2 q_{22} + \lambda_3 q_{12} \\ 1 & 1 & 2 & \lambda_1 q_{11} + \lambda_2 q_{11} + \lambda_3 q_{21} & \lambda_1 q_{12} + \lambda_2 q_{12} + \lambda_3 q_{22} \\ 2 & 1 & 2 & \lambda_1 q_{21} + \lambda_2 q_{11} + \lambda_3 q_{21} & \lambda_1 q_{22} + \lambda_2 q_{12} + \lambda_3 q_{22} \\ 1 & 2 & 2 & \lambda_1 q_{11} + \lambda_2 q_{21} + \lambda_3 q_{21} & \lambda_1 q_{12} + \lambda_2 q_{22} + \lambda_3 q_{22} \\ 2 & 2 & 2 & q_{21} & q_{22} \end{array}.$$

Since V_2, V_3 and V_4 have their last three rows equal, q_{11} cannot take a value too different from q_{21}, and q_{12} cannot take a value too different from q_{22}. This implies that the resulting MTD is not good at representing situations in which a column contains probabilities that are very different from one another. Thus the MTD-VLMC does not perform well for sequences generated using V_2, but it achieves better results for sequences generated using V_3 and V_4.

6.2 Other Models for Discrete-Valued Time Series

Raftery (1985b) proposed building a log-linear model using the same principle used for the MTD model, that is, by ignoring the interactions between lags. Raftery (1985b) also proposed a model inspired by the autoregressive moving-average (ARMA) model. The basic MTD model of (12) is modified as

$$\hat{\chi}_t' = \sum_{g=1}^{\ell} \lambda_g \chi_{t-g}' Q + \sum_{h=1}^{k} \mu_h \xi_{t-h}' Q + \mu_0 \xi' t,$$

where

$$\sum_{g=1}^{\ell} \lambda_g + \sum_{h=1}^{k} \mu_h + \mu_0 = 1$$

and $\xi_t = (\xi_t(1), \ldots, \xi_t(m))'$ is an indicator vector with $\xi_t(i) = 1$ if $Y_t = i$ and zero otherwise, where the Y_t are independent random variables with distribution π, $\pi' Q = \pi'$. The equilibrium distribution of X_t is π. This model is similar to the standard ARMA(ℓ, k) model, but it is no longer Markovian.

Liang and Zeger (1986) and Zeger and Liang (1986) developed the generalized estimating equation method for the estimation of GLM models in the case of longitudinal data with correlation among lags. Zeger and Qaqish (1988) introduced the class of Markov regression models. These observation-driven models are a generalization of a model presented previously by Cox (1981). They are very flexible and they can also incorporate covariates. In the binary case, the resulting models are similar to logistic regressions. Fahrmeir and Kaufmann (1987) described a similar model for the multinomial case. More details about this class of models can be found in Fahrmeir and Tutz (1994), Diggle, Liang and Zeger (1996), Laird (1996) and MacDonald and Zucchini (1997).

6.3 High-Order Hidden Markov Models

Hidden Markov models often use a first-order Markov chain to represent the transition process between hidden states. However, it seems interesting to be able also to use high-order dependencies. To avoid an excessive number of parameters, Schimert (1992) proposed replacing the higher order Markov chain by a MTD model in the context of speech recognition.

The same principle is also used in the double chain Markov model (DCMM) developed by Berchtold (1999, 2002). This model is based on the superposition of two Markov chains. A non-homogeneous observed process is described by a finite set of M transition matrices. At each time t, the choice of the active matrix is governed by a hidden homogeneous Markov chain. Both the hidden and visible transition matrices can be of order greater than 1, which implies a large number of parameters. This problem is solved through the use of a MTD model for each of these matrices.

6.4 Covariates

One possible way to incorporate covariates is to consider the observed data to be nonhomogeneous and to apply a model such as the DCMM of Section 6.3. In

this case, each set of external conditions leads to a different transition matrix between observed events. Alternatively, the covariates can be incorporated into the MTD model through the addition of a supplementary term or they can directly modify the basic MTD model. See Berchtold and Raftery (1999) for more details.

ACKNOWLEDGMENTS

This research was supported by Office of Naval Research Grant N00014-96-1-0192. André Berchtold was also supported by a grant from the Swiss National Science Foundation.

REFERENCES

ADKE, S. R. and DESHMUKH, S. R. (1988). Limit distribution of a high-order Markov chain. *J. Roy. Statist. Soc. Ser. B* **50** 105–108.

AITKIN, M., ANDERSON, D., FRANCIS, B. and HINDE, J. (1989). *Statistical Modelling in GLIM*, Chap. 6. Clarendon, Oxford.

BAUM, L. E. (1971). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Inequalities. III. Proceedings of a Symposium*. Academic Press, New York.

BAXTER, R. J. (1982). *Exactly Solved Models in Statistical Mechanics*. Academic Press, London.

BERCHTOLD, A. (1995). Autoregressive modelling of Markov chains. In *Proc. 10th International Workshop on Statistical Modelling* 19–26. Springer, New York.

BERCHTOLD, A. (1996). Modélisation autorégressive des chaînes de Markov: Utilisation d'une matrice différente pour chaque retard. *Rev. Statist. Appl.* **44** 5–25.

BERCHTOLD, A. (1997). Swiss health insurance system: Mobility and costs. *Health and System Science* **1** 291–306.

BERCHTOLD, A. (1998). *Chaînes de Markov et Modèles de Transition*: *Applications aux Sciences Sociales*. HERMES, Paris.

BERCHTOLD, A. (1999). The double chain Markov model. *Commun. Statist. Theory Methods* **28** 2569–2589.

BERCHTOLD, A. (2001). Estimation in the mixture transition distribution model. *J. Time Ser. Anal.* **22** 379–397.

BERCHTOLD, A. (2002). High-order extensions of the double chain Markov model. *Stoch. Models* **18** 193–227.

BERCHTOLD, A. and RAFTERY, A. E. (1999). The mixture transition distribution (MTD) model for high-order Markov chains and non-Gaussian time series. Technical Report 360, Dept. Statistics, Univ. Washington. Available at www.stat.washington.edu/www/research/reports/1999/tr360.ps.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

BESAG, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

BESAG, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64** 616–618.

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Reply to comments. *Ann. Inst. Statist. Math.* **43** 49–59.

BILLINGSLEY, P. (1961). *Statistical Inference for Markov Processes*. Univ. Chicago Press.

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

BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econometrics* **31** 307–327.

BOLLERSLEV, T., CHOU, R. Y. and KRONER, K. F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. *J. Econometrics* **52** 5–59.

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

BÜHLMANN, P. and WYNER, A. J. (1999). Variable length Markov chains. *Ann. Statist.* **27** 480–513.

COX, D. R. (1981). Statistical analysis of time series: Some recent developments (with discussion). *Scand. J. Statist.* **8** 93–115.

CRAIG, P. (1989). Time series analysis of directional data. Ph.D. thesis, Dept. Statistics, Trinity College, Dublin.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DIGGLE, P. J., LIANG, K.-Y. and ZEGER, S. L. (1996). *Analysis of Longitudinal Data*. Clarendon, Oxford.

DOMB, C. and POTTS, R. B. (1951). Order–disorder statistics. IV. A two-dimensional model with first and second interactions. *Proc. Roy. Soc. London Ser. A* **210** 125–141.

ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50** 987–1007.

FAHRMEIR, L. and KAUFMANN, H. (1987). Regression models for non-stationary categorical time series. *J. Time Ser. Anal.* **8** 147–160.

FAHRMEIR, L. and TUTZ, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York.

FRANCIS, B. et al. (1993). *The GLIM System*: *Release 4 Manual* (B. Francis, M. Green and C. Payne, eds.). Clarendon, Oxford.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

GRIMMETT, G. R. (1973). A theorem about random fields. *Bull. London Math. Soc.* **5** 81–84.

HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press.

HANEY, D. J. (1993). Methods for analyzing discrete-time, finite state Markov chains. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

HEALY, M. J. R. (1988). *GLIM*: *An Introduction*. Clarendon, Oxford.

HECKERMAN, D. J., CHICKERING, D. M., MEEK, C., ROUNTHWAITE, R. and KADIE, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *J. Machine Learning Research* **1** 49–75.

HOLGATE, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika* **51** 241–245.

HOLLAND, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Univ. Michigan Press, Ann Arbor.

HOSKING, J. R. M. (1981). Fractional differencing. *Biometrika* **68** 165–176.

JACOBS, P. A. and LEWIS, P. A. W. (1978a). Discrete time series generated by mixtures. I. Correlational and runs properties. *J. Roy. Statist. Soc. Ser. B* **40** 94–105.

JACOBS, P. A. and LEWIS, P. A. W. (1978b). Discrete time series generated by mixtures. II. Asymptotic properties. *J. Roy. Statist. Soc. Ser. B* **40** 222–228.

JACOBS, P. A. and LEWIS, P. A. W. (1978c). Discrete time series generated by mixtures. III. Autoregressive processes. Technical Report NPS 55-78-022, Naval Postgraduate School.

JACOBS, P. A. and LEWIS, P. A. W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Ser. Anal.* **4** 18–36.

JOHNSON, N. L. and KOTZ, S. (1969). *Distributions in Statistics*: *Discrete Distributions*. Houghton Mifflin, Boston.

KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

KASS, R. E. and RAFTERY, A.E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

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

KATZ, L. and PROCTOR, C. H. (1959). The concept of configuration of interpersonal relations in a group as a time-dependent stochastic process. *Psychometrika* **24** 317–327.

KEMENY, J. G. and SNELL, J. L. (1976). *Finite Markov Chains*. Springer, New York.

KEMENY, J. G., SNELL, J. L. and KNAPP, A. W. (1976). *Denumerable Markov Chains*. Springer, New York.

KINDERMAN, R. and SNELL, J. L. (1980). *Markov Random Fields and their Applications*. Amer. Math. Soc. Providence, RI.

KWOK, M. (1988). Some results on higher-order Markov chain models. Ph.D. thesis, Univ. Hong Kong.

LAIRD, N. M. (1996). Longitudinal panel data: An overview of current methodology. In *Time Series Models in Econometrics, Finance and Other Fields* (D. R. Cox, D. V. Hinkley and O. E. Barndorff-Nielsen, eds.). Chapman and Hall, London.

LE, N. D., MARTIN, R. D. and RAFTERY, A. E. (1996). Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *J. Amer. Statist. Assoc.* **91** 1504–1515.

LIANG K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

LLOYD, E. H. (1977). Reservoirs with seasonally varying Markovian inflows and their first passage times. Research Report RR-77-4, International Institute for Applied Systems Analysis, Laxenburg, Austria.

LOGAN, J. A (1981). A structural model of the higher-order Markov process incorporating reversion effects. *J. Math. Sociol.* **8** 75–89.

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

MARTIN, R. D. and RAFTERY, A. E. (1987). Robustness, computation, and non-Euclidean models. Comment on "Non-Gaussian state-space modelling of nonstationary time series," by G. Kitagawa. *J. Amer. Statist. Assoc.* **82** 1044–1050.

MCLACHLAN, G. J. and KRISHNAN, T. (1996). *The EM Algorithm and Extensions*. Wiley, New York.

MEHRAN, F. (1989a). Longitudinal analysis of employment and unemployment based on matched rotation samples. Report, International Labour Office, Bureau of Statistics, Geneva.

MEHRAN, F. (1989b). Analysis of discrete longitudinal data: Infinite-lag Markov models. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.) 533–541. North-Holland, Amsterdam.

PEGRAM, G. G. S. (1980). An autoregressive model for multilag Markov chains. *J. Appl. Probab.* **17** 350–362.

RAFTERY, A. E. (1985a). A model for high-order Markov chains. *J. Roy. Statist. Soc. Ser. B* **47** 528–539.

RAFTERY, A. E. (1985b). A new model for discrete-valued time series: Autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni* **3–4** 149–162.

RAFTERY, A. E. (1993). Change point and change curve modeling in stochastic processes and spatial statistics. *J. Appl. Statist. Sci.* **1** 403–423.

RAFTERY, A. E. and BANFIELD, J. D. (1991). Stopping the Gibbs sampler, the use of morphology and other issues in spatial statistics. *Ann. Inst. Statist. Math.* **43** 32–43.

RAFTERY, A. E. and TAVARÉ, S. (1994). Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model. *Appl. Statist.* **43** 179–199.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

SCHIMERT, J. (1992). A high order hidden Markov model. Ph.D. dissertation, Univ. Washington.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SUBBA RAO, T. and GABR, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models. Lecture Notes in Statist.* **24**. Springer, New York.

THEIL, H. (1971). On the estimation of relationships involving qualitative variables. *American J. Sociology* **76** 103–154.

WEINBERGER, M. J., RISSANEN, J. J. and FEDER, M. (1995). A universal finite memory source. *IEEE Trans. Inform. Theory* **41** 643–652.

WONG, C. S. and LI, W. K. (2000). On a mixture autoregression model. *J. Roy. Statist. Soc. Ser. B* **62** 95–115.

ZEGER, S. L. and LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

ZEGER, S. L. and QAQISH, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44** 1019–1031.