Estimation and Modelling Repeated Patterns in High Order Markov Chains with the Mixture Transition Distribution Model

Adrian Raftery; Simon Tavare

Applied Statistics, Vol. 43, No. 1 (1994), 179-199.

Stable URL:
http://links.jstor.org/sici?sici=0035-9254%281994%2943%3A1%3C179%3AEAMRPI%3E2.0.CO%3B2-B

Applied Statistics

Appl. Statist. (1994)
43, *No.* 1, *pp.* 179–199

Estimation and Modelling Repeated Patterns in High Order Markov Chains with the Mixture Transition Distribution Model

By ADRIAN RAFTERY

University of Washington, Seattle, USA

and SIMON TAVARÉ†

University of Southern California, Los Angeles, USA

[Received June 1991. Revised September 1992]

SUMMARY

The mixture transition distribution (MTD) model was introduced by Raftery as a parsimonious model for high order Markov chains. It is flexible, can represent a wide range of dependence patterns, can be physically motivated, fits data well and is in several ways a discrete-valued analogue for the class of autoregressive time series models. However, estimation has presented difficulties because the parameter space is highly non-convex, being defined by a large number of non-linear constraints. Here we propose a computational algorithm for maximum likelihood estimation which is based on a way of reducing the large number of constraints. This also allows more structured versions of the model, e.g. those involving structural zeros, to be fitted quite easily. A way of fitting the model by using GLIM is also discussed. The algorithm is applied to a sequence of wind directions, and also to two sequences of deoxyribonucleic acid bases from introns from mouse genes. In each case, the MTD model fits better than the conventional Markov chain model, and for the wind data it provides superior out-of-sample predictions. A modification of the model to represent repeated patterns is proposed and a very parsimonious version of this modified model is successfully applied to data representing bird songs.

Keywords: Angular data; Bird songs; Deoxyribonucleic acid; Discrete time series; Optimization; Wind directions

1. Introduction

There are many examples in which we would like to fit high order Markovian models to discrete data. However, in the conventional parameterization of such processes the number of parameters increases geometrically with the order, so that parsimony is effectively lost. In this paper, we describe some computational algorithms for fitting a parsimonious autoregressive-like Markov model known as the *mixture transition distribution* (MTD) model, and we illustrate its use with some examples.

In Section 3 we analyse a data set that includes over 77 000 hourly observations of wind directions at a meteorological station at Roche's Point, Ireland; see Raftery

†*Address for correspondence*: Departments of Mathematics and Biological Sciences, University of Southern California, 1042 W 36th Place, Los Angeles, CA 90089-1113, USA.

 0035-9254/94/43179

et al. (1982) and Haslett and Raftery (1989). One aim here is to predict wind speeds and directions to control the wind turbine generators that make up a wind farm, and to manage the electric power supply. Wind turbines should be oriented so that they derive the most energy from the wind, so that their current best orientation is a function of future as well as current wind direction. Predicting output from a variable energy source such as wind is important so that the need for power from other, more stable, sources such as oil can be anticipated.

Our second example, discussed in Section 4, is from the area of biomolecular sequence comparison. This field has provided statisticians with a wealth of novel problems. As an example, descriptive statistics on deoxyribonucleic acid (DNA) composition have proved useful in the search for coding regions and introns, the statistical assessment of sequence similarity and the analysis of repeated motifs that may be of biological significance. Similarly, a statistical analysis of protein sequences of known three-dimensional structures has been used to infer potential folding patterns of other proteins. It is not our purpose here to describe these areas in any detail; rather, we refer the reader to the recent review article by Curnow and Kirkwood (1989) and the books edited by Waterman (1988) and Doolittle (1990) for good introductions to this field. We shall focus on just one example from the area of DNA sequence analysis to which the MTD model might be applied.

Finally, in Section 5, we analyse some bird song data which are dominated by patterns that are repeated many times. Such data are used to study questions such as memory, responsiveness and imitation in birds. They may also be a template for other problems where complex repeating patterns are embedded in noise, such as speech recognition and the analysis of DNA coding regions. Here we find that a simple and parsimonious modification of the MTD model fits well.

The MTD model was introduced by Raftery (1985a, b) and is defined as follows. Let $\{X_t: t = 1, 2, \ldots\}$ be a time homogeneous lth-order Markov chain on a finite set of m states (here labelled $1, 2 \ldots, m$), and let the transition probabilities be

$$p(i_0 \mid i_1, \ldots, i_l) = P(X_{t+l} = i_0 \mid X_{t+l-1} = i_1, \ldots, X_t = i_l), \quad t = 1, 2, \ldots. \tag{1}$$

There are $(m-1)m^l$ independent parameters in equation (1). For $l > 1$, the MTD model provides a useful parameter reduction in equation (1) by supposing that

$$p(i_0 \mid i_1, \ldots, i_l) = \sum_{j=1}^{l} \lambda_j q(i_0 \mid i_j), \tag{2}$$

where $Q = \{q(i|j)\}$ is a *column* stochastic matrix satisfying

$$q(i|j) \geqslant 0 \quad \text{and} \quad \sum_{r=1}^{m} q(r|j) = 1, \qquad j = 1, \ldots, m, \tag{3}$$

and

$$\lambda_1 + \ldots + \lambda_l = 1. \tag{4}$$

Note that the number of independent parameters is now $m(m-1) + l - 1$, increasing only linearly in l. For example, when there are $m = 4$ states, the number of parameters for a second-order ($l = 2$) chain is 13 in the MTD model as against 48 in the usual second-order Markov chain model. To ensure that the transition probabilities are properly defined, we also require

$$\sum_{j=1}^{l} \lambda_j q(i_0 \mid i_j) \geqslant 0 \qquad \text{for all } i_0, \ldots, i_l. \tag{5}$$

Notice that when conditions (3)–(5) are satisfied $0 \leqslant p(i_0 \mid i_1, \ldots, i_l) \leqslant 1$ for all i_0, $i_1, \ldots, i_l$.

The MTD model is so called because the conditional probabilities in equation (2) are linear combinations of contributions from the past. It is analogous to the autoregressive AR(l) model in that one extra parameter is added to the model for each extra lag and that the lagged bivariate distributions satisfy a system of matrix equations that are similar to the Yule–Walker equations.

In some situations it has a direct physical interpretation in terms of the probability of returning to past states, or states close to them. For example, in modelling employment histories, someone who changed from job A to job B at the last transition may be more likely to revert to job A at the next transition than someone who had spent a long time in job B. Specifically, suppose that at the next time point $t + 1$ a person has probability λ_1 of being in a job like her present job and probability λ_2 of being in a job like the job that she occupied at the previous time point $t - 1$. Thus, if she changed job between times $t - 1$ and t, she has a probability λ_2 of reverting to her previous job or a job like it. Here the word 'like' has a stochastic meaning defined by the matrix Q: if a person leaves job j to move to one like it, the conditional probability that the new job is i is $q(i|j)$ $(i = 1, \ldots, m)$. Then the person's job history is precisely a trajectory of the stochastic process defined by equation (2) with $l = 2$. If a person has a non-zero probability of reverting to jobs like those held up to l time periods ago, we have model (2) with general l. Raftery (1985a) analysed a set of job histories of physicists and showed them to be well fitted by this model.

We have found the MTD model useful in practice. From an algorithmic point of view, there are really two special cases of the MTD model, these being determined by what is assumed in addition to conditions (3) and (4) about the $\{\lambda_i\}$. With the positivity assumption

$$\lambda_i \geqslant 0, \qquad i = 1, \ldots, l, \tag{6}$$

the inequality in expression (5) is automatically satisfied, and fitting presents few problems. This is the most important case in practice, and it has the added benefit that the λ-parameters may be readily interpreted as probabilities. The examples given in Raftery (1985a) satisfied assumption (6).

However, not all the data sets that we have analysed satisfy this condition. Without the positivity assumption, equation (5) comes into play in a crucial way, and computationally it becomes important to be able to reduce the large number of constraints that are operating there. In Section 2 we suggest how this might be done and in Section 4 we give an example.

2. Parameter Estimation

2.1. *Maximum Likelihood Estimation*

The parameters Q and $\{\lambda_i\}$ can be fitted by maximum likelihood by maximizing the log-likelihood

$$\log L = \sum n(i_0, i_1, \ldots, i_l) \log p(i_0 \mid i_1, \ldots, i_l), \tag{7}$$

where $n(i_0, i_1, \ldots, i_l)$ is the number of times that the sequence $i_l \to i_{l-1} \to \ldots \to i_0$ occurs in the data, $p(i_0 \mid i_1, \ldots, i_l)$ is given by equation (2) and the sum is over all $i_0, i_1, \ldots, i_l$ with $n(i_0, i_1, \ldots, i_l) > 0$. The maximization is subject to constraints (3), (4) and (6). In equation (7) we have ignored the contribution to the likelihood that comes from the initial distribution.

Although there are several numerical approaches that might be taken to this problem, we found that direct maximization of log-likelihood (7) was effective. We used the sequential quadratic programming algorithm implemented as routine E04UCF in mark 15 of Numerical Algorithms Group (1991). Although derivatives of the objective function and the constraint functions may be calculated, we found that approximating these by finite differences was effective.

One troublesome part of the algorithm involves the storage and recovery of the counts $n(i_0, i_1, \ldots, i_l)$. For models with high values of l or m, the number m^{l+1} of potential patterns can be extremely large. We proceeded by labelling a pattern $(i_0, i_1, \ldots, i_l)$ by the number

$$i = 1 + \sum_{j=0}^{l} (i_j - 1) m^{l+1-j}.$$

If the number of possible patterns was sufficiently small, we stored the whole (now one-dimensional) array of counts. However, the maximum number N, say, of patterns that can be observed in the data is a little less than the length of the observed time series, so in cases where m^{l+1} is a very large compared with N we calculated and stored the counts by using a simple hashing algorithm with a vector of length approximately $1.25N$.

The types of data that we have analysed have led to several useful additional features of the programs.

(a) Structural zeros in the Q-matrix may be handled directly. Example 3.2 in Raftery (1985a) is of this type.

(b) Predetermined λ-values may also be set to 0, corresponding to the omission of those terms.

(c) Fixed or random starts for the parameters in the iterative scheme are allowed. In the first instance, Q is estimated by the usual first-order transition matrix, and the λ_i are equal. In the second instance, random Q and λ_i are used, subject to constraints (3) and (4). This facility is particularly useful in the iterative algorithm for determining whether a local or potentially global maximum has been reached. Although we have no formal proof that a unique maximum exists, numerical evidence with some 20 data sets suggests that it does.

2.2. *Reducing Number of Constraints*

In the absence of the positivity constraints (6), the general MTD model must satisfy the $m^l(m-1)$ constraints in expression (5). For example, in the four-state second-order case, i.e. $m = 4$ and $l = 2$, the number of constraints is 48, so that the resulting constrained numerical optimization problem is computationally

demanding. The following result effectively reduces the number of constraints in expression (5) to m.

Proposition 1. Let $T = \Sigma_{i:\lambda_i \geqslant 0} \lambda_i$, and define $q_-(i) = \min_{1 \leqslant j \leqslant m} \{q(i|j)\}$ and $q_+(i) = \max_{1 \leqslant j \leqslant m} \{q(i|j)\}$. Then $\Sigma_{j=1}^{l} \lambda_j q(i|i_j) \geqslant 0$ for all $i, i_1, \ldots, i_l$, if and only if

$$Tq_-(i) + (1 - T)\, q_+(i) \geqslant 0 \text{ for all } i. \tag{8}$$

Proof. If inequality (8) holds, then

$$\begin{aligned}\sum_{j=1}^{l} \lambda_j q(i|i_j) &= \sum_{j:\lambda_j \geqslant 0} \lambda_j q(i|i_j) + \sum_{j:\lambda_j < 0} \lambda_j q(i|i_j) \\ &\geqslant \sum_{j:\lambda_j \geqslant 0} \lambda_j q_-(i) + \sum_{j:\lambda_j < 0} \lambda_j q_+(i) \\ &= Tq_-(i) + (1 - T)\, q_+(i) \\ &\geqslant 0.\end{aligned}$$

Conversely, assume that condition (5) holds, and suppose that $q_-(i) = q(i|p_0)$ and $q_+(i) = q(i|p_1)$. Then

$$\begin{aligned}Tq_-(i) + (1 - T)\, q_+(i) &= \sum_{j:\lambda_j \geqslant 0} \lambda_j q(i|p_0) + \sum_{j:\lambda_j < 0} \lambda_j q(i|p_1) \\ &\geqslant 0,\end{aligned}$$

which completes the proof.

The likelihood function in equation (7) can now be maximized under the constraints given in inequality (8). However, the constraint function is highly non-differentiable, and this leads to problems with the use of routine E04UCF. Although the algorithm converges close to a solution quite rapidly, it then cycles in a neighbourhood of the local optimum. As a simpler, if somewhat *ad hoc*, procedure we run the algorithm with *no* constraints on the λ-terms and then use proposition 1 to check that the solution does satisfy the constraints (5). The fits with negative λ-values described in Section 4 were found in this way.

2.3. *Minimum χ^2 Estimation*

As an alternative to maximum likelihood, we have also used minimum χ^2 estimation. The aim is to find Q and $\{\lambda_i\}$ that minimize

$$X^2 = \sum \frac{\{n(i_0, \ldots, i_l) - e(i_0, \ldots, i_l)\}^2}{e(i_0, \ldots, i_l)},$$

where

$$e(i_0, \ldots, i_l) = n(+, i_1, \ldots, i_l)\, p(i_0 | i_1, \ldots; i_l),$$

and + denotes summation over that index. The sum is over all $n(i_0, i_1, \ldots, i_l)$ for which $n(+, i_1, \ldots, i_l) > 0$, and the constraints (3), (4) and (8) apply. This is a useful alternative, since the fitted counts e from the optimization are natural

candidates as measures of goodness of fit of the model. Kwok (1988) and Li and Kwok (1989) have shown that in some special cases of the MTD model the minimum χ^2 estimator has lower bias than the maximum likelihood estimator but about the same variance, and hence lower overall mean-squared error.

The asymptotic theory of X^2 when the parameters have been estimated by maximum likelihood is given in Billingsley (1961). He shows that, under the assumption that the process is described by an lth-order Markov chain, X^2 has asymptotically as $n \to \infty$ a χ^2-distribution.

We also experimented with different numerical algorithms for this problem, essentially based on knowledge of derivatives for the constraint functions. Once more the direct approach seems the easiest, using algorithm E04UCF again.

2.4. *GLIM Analysis of Two-state Models*

When $m = 2$, the MTD model may be fitted by using an iterative procedure in GLIM (Baker, 1986). Focus for the moment on the case $l = 2$, and write $\lambda_1 = \lambda$, $\lambda_2 = 1 - \lambda$. Log-likelihood (7) may be written

$$\log L = \sum_{i_1, i_2} \left\{ \sum_{i=1}^{2} n(i, i_1, i_2) \log p(i | i_1, i_2) \right\} . \tag{9}$$

For each i_1 and i_2, the inner term in equation (9) is (essentially) a binomial log-likelihood for $n(+, i_1, i_2)$ trials and success probability $p(1 | i_1, i_2)$, where

$$p(1 | i_1, i_2) = \begin{cases} q(1|1) & i_1 = 1, i_2 = 1, \\ \lambda q(1|1) + (1 - \lambda) q(1|2) & i_1 = 1, i_2 = 2, \\ (1 - \lambda) q(1|1) + \lambda q(1|2) & i_1 = 2, i_2 = 1, \\ q(1|2) & i_1 = 2, i_2 = 2. \end{cases} \tag{10}$$

If λ is known, then equations (10) show that the $p(1 | i_1, i_2)$ are linear in the parameters $q(1|1)$ and $q(1|2)$, $q(1|1)$ being the coefficient of the covariate $\mathbf{x}_1^{\mathrm{T}} = (1, \lambda, 1 - \lambda, 0)$ and $q(1|2)$ the coefficient of the covariate $\mathbf{x}_2^{\mathrm{T}} = (0, 1 - \lambda, \lambda, 1)$. Thus $q(1|1)$ and $q(1|2)$ (and so Q) may be estimated by using binomial error, identity link, no intercept and covariates $\mathbf{x}_1$ and $\mathbf{x}_2$.

However, if $q(1|1)$ and $q(1|2)$ are assumed known, then equations (10) show that $p(1 | i_1, i_2)$ is linear in λ; λ is the coefficient of the covariate $\mathbf{x}_3^{\mathrm{T}} = (0, q(1|1) - q(1|2), q(1|2) - q(1|1), 0)$ and the offset is $\mathbf{x}_4^{\mathrm{T}} = (q(1|1), q(1|2), q(1|1), q(1|2))$. Thus λ may be estimated by using binomial error, identity link, no intercept, covariate $\mathbf{x}_3$ and offset $\mathbf{x}_4$. This leads to a simple recursive scheme for estimating the parameters, reminiscent of the iterative algorithms used in survival analysis; see Aitkin *et al.* (1989).

The generalization to $l > 2$ is almost immediate from the form of equations (10). The number of covariates for the first stage remains 2, the elements of $\mathbf{x}_1$ being replaced by $\Sigma_{j:i_j = 1} \lambda_j$. For the second stage the number of covariates is $l - 1$. It does not, however, seem to be simple to generalize this scheme to the case $m > 2$.

2.5. *Model Comparison*

To compare the rival, non-nested, models in the examples that follow, we would ideally like to compute the posterior probability of each model under a range of

plausible prior distributions for the parameters. The use of successive significance tests seems less satisfactory because many of the comparisons involve non-nested models and because the use of multiple tests makes the properties of the overall procedure difficult to assess.

Here we use the approximate result that if we are comparing two models M_0 and M_1 then the Bayes factor, or ratio of posterior to prior odds, B_{01}, for M_0 against M_1 satisfies

$$-2\log B_{01} \sim \mathrm{BIC}_0 - \mathrm{BIC}_1. \tag{11}$$

In approximation (11) $\mathrm{BIC}_i = -2\log L_i + k_i \log n$, where L_i is the maximized likelihood and k_i is the number of independent parameters in the model M_i ($i = 0, 1$). This holds quite generally for models which satisfy regularity conditions sufficient for the maximum likelihood estimator to be asymptotically normally distributed (e.g. Raftery (1988)). The MTD model (2) does satisfy such conditions provided that the parameters lie in the interior of the parameter space, i.e. that the inequalities in equation (5) are all strict. We assume this to be the case here. Result (11) has been formally established for independent exponential family observations by Schwarz (1978), for the usual Markov chain by Katz (1981) and for log-linear models of contingency tables by Raftery (1986a).

If M_0 is always some base-line model such as the independence model and expression (11) is calculated for each model of interest M_1, then the resulting Bayes factors yield the approximate posterior probabilities of each of the models of interest (Raftery, 1988). The rules of thumb of Jeffreys (1961), appendix B, suggest that such a comparison should not be regarded as decisively favouring a larger model over a smaller nested model unless the difference in BIC values is at least about 10.

Model comparisons based on posterior probabilities can yield results that are different from those based on significance tests. This is especially so with large samples, including some that we analyse here. In such cases significance tests at fixed significance levels often reject null hypotheses more easily; an example with $n \approx 110\,000$ was discussed by Raftery (1986b). This is related to the 'conflict between significance and P values' discussed by Berger and Sellke (1987). Alternatively, basing model comparisons on Bayes factors may be viewed as an automatic, decision theoretic, way of setting significance levels to balance power and significance.

Our code produces Pearson residuals which can be used to suggest ways in which the model could be improved. New models suggested by such a process can be compared with the other models under consideration also using approximate Bayes factors.

3. Wind Direction Data

We return to the observations on wind directions at Roche's Point, Ireland. The data were recorded from 1.00 a.m. on January 1st, 1961, and ran for almost 9 years. There are 77 155 observations in all. The original data were recorded as 0 for no wind, and then in 10° units from due north, for a total of 37 states.

Fig. 1 provides a histogram of the distribution of wind directions for all 9 years combined, together with separate histograms for each of the 8 complete years in the data set. Broadly, these annual histograms are rather similar and show natural

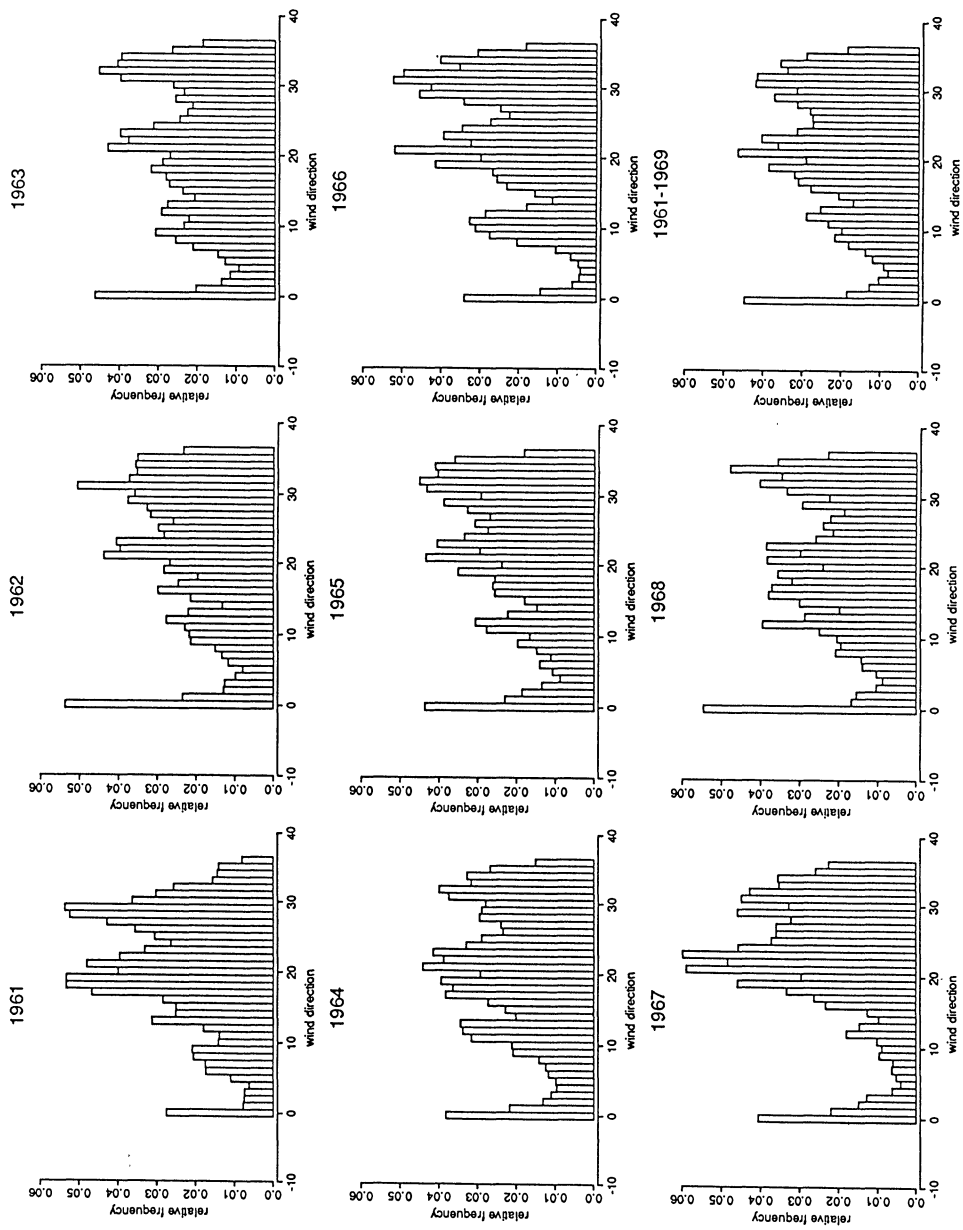

Fig. 1. Histograms of wind directions in Roche's Point, Ireland, 1961–69: analysis by year (wind directions are measured in units of 10° from north; 0 indicates no wind)

modes in the data that are preserved from year to year. On the basis of these results, we chose to recode the data into five categories: 0, 6–14, 15–23, 24–32 and 33–5; these are labelled 1–5 respectively in what follows.

As might have been expected, there are some inhomogeneities in the distribution of wind directions which are revealed when the data are analysed in separate months. See Fig. 2.

These distributions are rather similar for the months November–April. We therefore chose the months November–April as a period in which wind directions might be modelled by a stationary MTD model. The data analysed here come from the period November 1961–April 1962, providing a total of 4344 consecutive hourly observations.

Table 1 gives the BIC values for the full Markov model and the MTD model. The order is estimated to be 7. The estimated parameters are $\hat{\lambda}_1 = 0.591$, $\hat{\lambda}_2 = 0.237$, $\hat{\lambda}_3 = 0.076$, $\hat{\lambda}_4 = 0.018$, $\hat{\lambda}_5 = 0.024$, $\hat{\lambda}_6 = 0.024$ and $\hat{\lambda}_7 = 0.031$, whereas

$$Q = \begin{pmatrix} 0.65 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.09 & 0.95 & 0.01 & 0.00 & 0.03 \\ 0.14 & 0.02 & 0.92 & 0.04 & 0.00 \\ 0.05 & 0.00 & 0.06 & 0.92 & 0.04 \\ 0.08 & 0.03 & 0.00 & 0.03 & 0.91 \end{pmatrix}$$

The estimated $\hat{\lambda}_j$ are positive and decline with lag, indicating that the most recent observations are the most important, and that the current observation tends to be close to the immediately preceding observations. The $\hat{Q}$-matrix indicates the process to be smooth, with the probability of staying in the same state being 0.91 or greater whenever there is any wind, and the probability that the direction changes by more than one state being very small.

We note that $\hat{\lambda}_7$ is larger than $\hat{\lambda}_4$, $\hat{\lambda}_5$ and $\hat{\lambda}_6$, probably because the λ_7-term is capturing the small residual dependence on X_{t-8}, X_{t-9}, ..., as well as the dependence on X_{t-7} itself. This suggested that we fit another MTD(7) model, with the constraint that $\lambda_4 = \lambda_5 = \lambda_6 = 0$. As we discussed in Section 2.2, this is easy with our algorithm.

TABLE 1
Wind direction data from Roche's Point, Ireland†

Order l	*No. of parameters, k*	*BIC (full model)*	*No. of parameters, k*	*BIC (MTD model)*
0	4	12716.5		
1	20	5085.7	20	5085.7
2	100	5198.8	21	4646.4
3	500	8243.3	22	4569.5
4	2500	24674.7	23	4557.5
5	–	–	24	4544.7
6	–	–	25	4539.8
7	–	–	26	4538.8
8	–	–	27	4540.8
9	–	–	28	4540.4
10	–	–	29	4545.4

†n = 4325 observations starting at position 20 in the sequence.

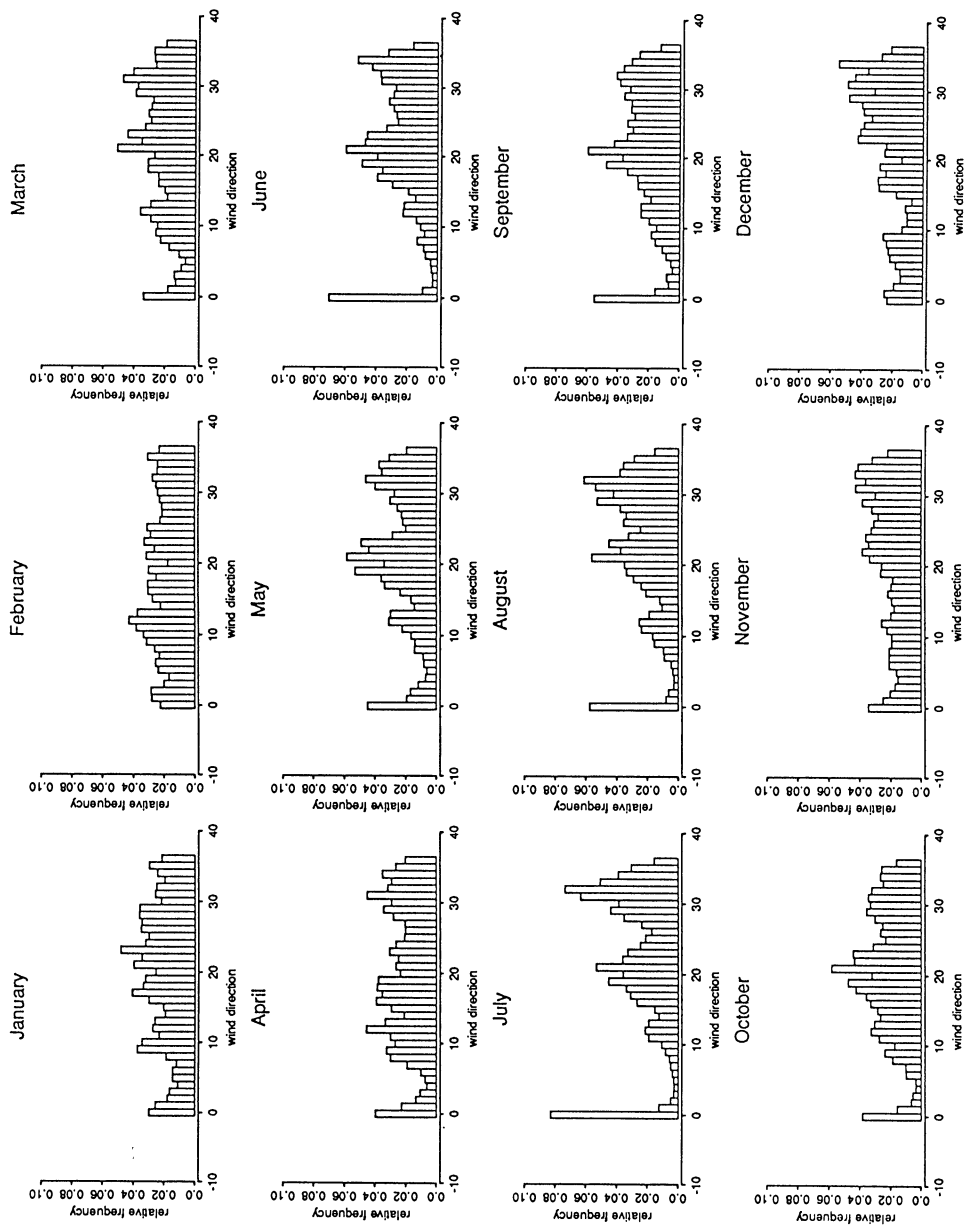

Fig. 2. Histograms of wind directions in Roche's Point, Ireland, 1961–69: analysis by month (wind directions are measured in units of 10° from north; 0 indicates no wind)

TABLE 2
Prediction results for the wind data†

i	*Model*	D_i
1	Independent trials (equal frequencies)	1609.4
2	Independent trials (Markov, order 0)	1779.4
3	Markov, order 1	699.1
4	MTD(7) ($\lambda_4 = \lambda_5 = \lambda_6 = 0$)	658.0

†Results based on fitting models to the first 3844 observations, and using the last 500 for prediction.

The resulting BIC value was 4530.4, making this quite clearly the best model considered. The $\hat{Q}$ -matrix was almost unchanged, whereas $\hat{\lambda}_1 = 0.598$, $\hat{\lambda}_2 = 0.245$, $\hat{\lambda}_3 = 0.001$ and $\hat{\lambda}_7 = 0.057$. This is not very different from the full MTD(7) model, but it seems to summarize the dependence in a more parsimonious way.

To assess the predictive power of the MTD model, we used a variant of the hold-out procedure. We refitted the models to the first 3844 data points and used the last 500 observations for comparing models. See Dawid (1984, 1986) for further discussion of this general approach to comparing models. We used the fitted parameters for model i to compute the predictive probability P_i of observing the last 500 data points, conditional on the earlier data. To assess the out-of-sample predictive performance of each model i, we used the logarithmic scoring rule of Good (1952) in the form $D_i \equiv -2 \log P_i$. This is on the same scale as the deviance in GLIM, and we refer to it as the 'predictive deviance'. Lower values of D_i correspond to better out-of-sample predictive performance. The results are shown in Table 2.

Our preferred model, the MTD(7) model with $\lambda_4 = \lambda_5 = \lambda_6 = 0$, is clearly better than the more conventional first-order Markov chain, outperforming it by 41 units of predictive deviance. Also, observe from the results for models 1 and 2 that it is not necessary for a more highly parameterized model to outperform a submodel in terms of predictive deviance, in contrast with the standard likelihood calculation.

4. Analysis of Intron Sequences

The statistical significance of repeated patterns in a DNA sequence must be measured against the background stochastic structure of the sequence itself. Among possible models for this structure are Markov chains, which might describe the DNA sequence in terms of its nucleotide composition (i.e. as a string of letters from a four-letter alphabet, {A, C, G, T}). There are several other alphabets of biological interest such as the purine–pyrimidine alphabet in which each base in the sequence is coded as either purine ({A,G}) or pyrimidine ({C,T}). For example, Blaisdell (1983) reported that, relative to a model of independent bases, non-coding sequences (such as introns) generally contain a shortage of runs of length 1 and 2 of purines and pyrimidines, and an excess of long runs of them; see also Karlin *et al.* (1988). In this section, we describe an exploratory analysis of two different DNA sequences from introns in certain mouse genes.

4.1. *Mouse T-cell Receptor α/δ-locus*

The first example is an analysis of part of the mouse T-cell receptor α/δ-locus (Wilson *et al.*, 1992; Koop *et al.*, 1992). This region is 94 647 bases in length. It comprises over 50 introns and 50 exons; the exons comprise just 6% of the sequence. The particular sequence that we have analysed is the intron before joining gene segment J50 (Koop *et al.*, 1992). It starts 5′ to exon 1 of Vδ5 and ends three bases before the recombination signal 5′ to J50. The sequence is 5778 bases in length.

A preliminary analysis shows that the sequence is clearly first order when analysed in the four-letter alphabet {A, G, C, T}; see Table 3. The MTD model provides no improvement on this fit. The estimated transition matrix $\hat{Q}$ is given in Table 4.

A Markov chain with transition matrix Q is (strongly) lumpable with respect to a partition $P_1, \ldots, P_r$ of the state space if, and only if, for $1 \leqslant (i, j) \leqslant r$, $\Sigma_{l \in P_j} q(l|k)$ has the same value for each $k \in P_i$ (Kemeny and Snell, 1960). The lumpability condition ensures that the lumped process is also Markovian, no matter

TABLE 3
Intron from mouse T-cell receptor α/δ-locus†

Order l	*No. of parameters, k*	*BIC (full model)*	*No. of parameters, k*	*BIC (MTD model)*
Alphabet A, C, G, T				
0	3	15 980.1		
1	12	15 549.3	12	15 549.3
2	48	15 756.9	13	15 552.4
3	192	16 828.7	14	15 561.1
Alphabet A/G, C, T				
0	2	12 042.0		
1	6	11 900.1	6	11 900.1
2	18	11 939.7	7	11 885.5
3	54	12 204.7	8	11 890.2
Alphabet A/G, C/T				
0	2	8005.6		
1	4	7881.8	4	7881.8
2	8	7850.5	3	7851.4
3	16	7878.2	4	7856.1

†n = 5769 bases, starting at position 10 in the sequence.

TABLE 4
Estimated Q-matrix for the Markov model

Next base	*Last base*			
	A	*G*	*C*	*T*
A	0.31	0.29	0.32	0.18
G	0.27	0.27	0.04	0.29
C	0.20	0.22	0.27	0.26
T	0.22	0.21	0.37	0.27

what the initial distribution of the original process might be. An examination of the matrix $\hat{Q}$ in Table 4 indicates that we might simplify the stochastic description of this intron by lumping the states A and G into a single state A/G that denotes purine. A formal test of this lumpability hypothesis may be found in Thomas and Barr (1977). The new alphabet is {A/G, C, T}. The BIC analysis of this new sequence is presented in Table 3. As would be expected, among the fully parameterized Markov models a first-order chain provides the most parsimonious description. Its estimated transition matrix is given in Table 5. However, it can be seen from Table 3 that an MTD(2) model provides a better description. In this case, $\hat{\lambda}_1 = 0.71$ and $\hat{\lambda}_2 = 0.29$, and the estimated transition matrix $\hat{Q}$ is given in Table 6.

Finally, we analyse the purine-pyrimidine alphabet {A/G, C/T}. From the previous discussion, it seems clear that the original sequence is not lumpable with respect to the purine-pyrimidine partition of the states. The purine-pyrimidine sequence may not be Markovian (or even homogeneous, unless we assume that the original chain was stationary). Fitting Markovian models to such data provides an exploratory approach to approximating the stochastic structure of a complicated process by simpler processes. We might expect this more complicated structure to be reflected in a higher estimated order of dependence in the Markovian approximation. This is indeed the case here, as the results in Table 3 verify. The purine-pyrimidine chain is approximated by a second-order Markov chain. The MTD model offers no further improvement in this case.

The discussion of lumpability provides an indication of the greatest extent to which the states of the chain can be aggregated without losing important structure. Here, this seems to be the three-state case analysed above. Thus, in a sense, the second-order MTD model for the three-state case provides the most parsimonious

TABLE 5
Estimated Q-matrix for the Markov model

Next base	*Last base* A/G	C	T
A/G	0.57	0.36	0.47
C	0.21	0.27	0.26
T	0.22	0.37	0.27

TABLE 6
Estimated Q-matrix for the MTD(2) model

Next base	*Last base* A/G	C	T
A/G	0.60	0.33	0.44
C	0.20	0.28	0.26
T	0.20	0.39	0.29

available representation of the data within the class of Markov chain models discussed.

4.2. *Mouse αA-crystallin Gene*

Avery (1987) examined the Markovian structure of introns from several other genes in mice, to determine whether certain short DNA sequences occurred more often than would be expected by chance. Here we shall analyse the introns from the mouse αA-crystallin gene, further details of which may be found in Avery (1987). The sequence analysed here comprises two introns, of total length 1307 bases, and is shown in Table 7.

Following the style of analysis of the previous example, we see first that the

TABLE 7
Data for introns from mouse αA-crystallin gene; see Avery (1987)†

3411314321	1221212433	4422422413	2432421313	3334244334	2443424131	3312132413	3314224231	3333212414	3122243432
4242242113	2112342221	3224242221	3224234122	2332131334	4333434321	2212143224	1432431134	2122212142	1223331443
3334233133	1331312211	2213222231	3222213222	3124322434	2444422422	2142133211	3441214332	3113321132	3432443122
3322124222	4331244424	3442443343	4333334433	3143422431	1322213131	3312333244	4244121312	3333213311	2244243413
2433211132	4321314333	3242244424	1324422243	1431243113	1113322121	4222422241	2221134313	2212134424	1311313243
2242224443	2442111131	1332422213	1233314221	3243211242	4133113211	1443222211	1432133242	4132133321	3311313133
3343331113	3244434344	1321324134	1341243433	1124344334	1141433121	3323133313	3242243312	3414331412	1221213124
4413422212	2132413221	3442221144	1322133211	1431422441	1431122232	2122214114	2112214112	2112213241	2244411421
1221324324	4114243441	1313221422	1112214112	2342121241	3122141142	1132112233	2211224341	1241321442	4324312112
2133213244	1432411342	1411322141	1224331242	4114213221	1332124212	4312214421	1221442412	4241222112	2242411424
1411221344	1324414424	1411421344	3424312221	3414122213	2114311423	2124322213	4224433342	4331334211	4343314244
1224343333	2213343121	3422213222	2133121212	2132444244	3324431331	1321313313	4312334433	4424314332	4313242442
2443422414	3431422132	2214213312	2141213243	4422222142	2444342424	3313243343	3422444333	3132422111	2343334443
3444213									

†To be read across the rows: A≡1, C≡2, G≡3, T≡4.

TABLE 8
Introns from mouse αA-crystallin gene†

Order l	*No. of parameters, k*	*BIC (full model)*	*No. of parameters, k*	*BIC (MTD model)*
Alphabet A, C, G, T				
0	3	3620.8		
1	12	3559.7	12	3559.7
2	48	3758.8	13	3566.1
3	192	4542.8	14	3572.8
Alphabet A/G, C, T				
0	2	2739.0		
1	6	2728.7	6	2728.7
2	18	2786.6	7	2722.7
3	54	2973.2	8	2729.4
Alphabet A/G, C/T				
0	1	1810.9		
1	2	1792.8	2	1792.8
2	4	1798.1	3	1791.3
3	8	1813.8	4	1797.1

†Two introns, n = 1302 bases, starting at position 6 in the sequence.

sequence of bases with alphabet {A, C, G, T} is clearly indicated to be of order 1; see Table 8. The estimated transition matrix $\hat{Q}$ is given by Table 9.

Notice that this transition matrix is qualitatively rather similar to the corresponding matrix for the T-cell receptor intron discussed in the previous section. In particular, this sequence is also (approximately) lumpable with respect to the partition {A/G, C, T}. The results are again consistent with the previous example, in that among the fully parameterized Markov models the first-order model provides the best description. However, a better description is provided by the MTD(2) model in which $\hat{\lambda}_1 = 2.46$ and $\hat{\lambda}_2 = -1.46$, and the estimated transition matrix $\hat{Q}$ is given in Table 10. Note that in this example the likelihood is maximized by some negative values of the λ_i. The constraints are checked by using the method outlined after proposition 1.

Finally, the analysis of the collapsed chain in its purine–pyrimidine alphabet is given in Table 8. The odds for the data being generated by a second-order MTD model as against a first-order Markov chain are about 2:1 by expression (11). This provides some evidence for the chain being of order 2, although in the words of Jeffreys (1961) it is 'not worth more than a bare mention'. (The standard likelihood ratio test statistic is 8.7 with 1 degree of freedom, and the corresponding P-value from the asymptotic χ^2-distribution is about 0.003. Thus the approximate Bayes factor and the approximate P-value both support the MTD(2) model but, as usual, the P-value suggests stronger evidence for the larger model.) The parameter estimates from the GLIM algorithm described in Section 2.4 are identical, and the minimum χ^2 estimates are essentially the same.

In this example, the structure of the purine–pyrimidine sequence is captured by the MTD model, rather than by the fully parameterized Markov model that was

TABLE 9
Estimated Q-matrix for the Markov model

Next base	*Last base* A	G	C	T
A	0.23	0.23	0.30	0.19
G	0.34	0.32	0.06	0.30
C	0.25	0.27	0.34	0.28
T	0.18	0.19	0.30	0.23

TABLE 10
Estimated Q-matrix for the MTD(2) model

Next base	*Last base* A/G	C	T
A/G	0.52	0.43	0.49
C	0.27	0.32	0.29
T	0.21	0.25	0.22

TABLE 11
Estimated Q-matrix for the MTD(2) model

Next base	*Last base* A/G	C/T
A/G	0.52	0.45
C/T	0.48	0.55

required for the earlier intron sequence. The parameters are $\hat{\lambda}_1 = 2.19$ and $\hat{\lambda}_2 = -1.19$, and the estimated transition matrix $\hat{Q}$ is given in Table 11.

4.3. *Comments*

In these examples the emphasis is on model fitting to find (or approximate) structure, rather than for prediction. We have seen that the MTD(2) model provides a good description of the two intron sequences when they are coded in the {A/G, C, T} alphabet. Some other sequence analysis examples in which the MTD model has been applied appear in Tavaré and Giddings (1988). Although Markov models are a useful first step in this context, their validity is often questionable because of possible inhomogeneities in the sequence. This inhomogeneity is particularly pronounced in coding regions (exons), where it is well known that the three codon positions exhibit markedly different behaviour. To analyse such regions, more sophisticated non-homogeneous Markov models may be required. Some of these are described, for example, by Borodovsky *et al.* (1986) and Watterson (1992). The models developed in the next section may also be useful for this.

5. Modelling Repeated Patterns: Song of Wood Pewee

The MTD model will not work for all data sets. Like the autoregressive models for continuous data, it is designed for situations where the next observation is 'like' or 'unlike' the previous observations, but where there are not strong interactions between the past and present observations in their effect on the next observation (i.e. the effects are additive). This will not be the case when the data are dominated by repeated patterns. An example of this is the song of the wood pewee described by Craig (1943), parts of which were reanalysed by Chatfield and Lemon (1970) and Bishop *et al.* (1975) by using conventional Markov chain methods. The wood pewee is one of the song birds of eastern North America and its morning twilight song has three distinct phrases, labelled 1, 2 and 3.

Here we reanalyse one of the records of the morning twilight song of length $n = 1327$, reproduced in Table 12. The data are dominated by the pattern 1312, which occurs 260 times. This pattern is of *length* 4, but it can be specified by four transitions of *order* 2, namely 2|13, 1|21, 3|12 and 1|31; we shall say that the pattern is of order 2. It is therefore not surprising that a full Markov chain model of order 2 fits much better than a full Markov chain model of any other order, or than any MTD model; see Table 13.

TABLE 12
Morning twilight song of the wood pewee, from Craig (1943)†

2222222211 2112112112 1121121131 2121121121 3121312121 3121312131 2131213121 2112121312 1312131213 1213121312
1312131213 1213121312 1312131213 1213121312 1311312131 2131131213 1213121312 1312131211 2131213121 2131213121
3121312131 2131213121 3121312121 1213121312 1312131213 1213112131 2131213121 3121312131 2131213121 3121312131
2131213121 3121312131 2131213121 2131213121 3121312131 2131213121 3121312113 1121213121 3121312131 2131213121
3121312131 2131213121 3121312131 2131212131 2131213131 2131231312 1312131213 1121312131 2131213121 2131213121
3123121312 1312131213 1213121312 1312121312 1312131213 1213121312 1312131213 1213121312 1312131213 1213121312
1312131213 1213121312 1312131213 1213121312 1312131213 1213121312 1312131213 1213121312 1312131213 1213121312
1312131213 1213121312 1312131212 1312131213 1213121312 1312131213 1213121312 1312131213 1213121213 1213121312
1312131213 1213121312 1312131213 1213121312 1312131213 1213121312 1312131213 1213121312 1312131212 1312131213
1213121312 1312131213 1213121312 1312131213 1213121312 1312131213 1212131213 1213131211 2131212131 2131211312
1312131213 1211312131 2131213121 1213121312 1312121312 1312131213 1212131211 2131121121 1321121131 2121312131
2131213121 2131213121 1213121312 1213121312 1312112131 1213121312 1131211211 2112112112 1112131211 2112131213
1211211213 1213121211 3121212112 1312121211 2311213121 1213121121 3121211121 1213121213 1212131211 2111121121
1211211211 2112112112 1112112

†To be read across the rows.

TABLE 13
Results for the wood pewee song data†

Model	*No. of parameters, k*	*BIC*
1, full model, order 0	2	2713.3
2, full model, order 1	6	1431.4
3, full model, order 2	18	866.6
4, full model, order 3	54	1096.1
5, MTD model, order 2	7	1338.9
6, MTD model, order 3	8	1343.6
7, 1312 pattern + MC(0)‡	3	997.5
8, 1312 pattern + MC(1)	7	1021.1
9, 1312 and 112 patterns + MC(−1)	4	827.0
10, 1312 and 112 patterns + MC(0)§	6	832.6
11, 1312 and 112 patterns + MC(1)	10	858.7

†n = 1327 phrases, starting from position 5; data in Table 12.
‡MC(k) denotes the full model of order k. The order 0 model refers to an independent sequence whereas the order −1 model refers to an independent and equiprobable sequence. Model 7 is defined by equation (12).
§Model 10 is defined by equation (13).

The idea underlying the MTD model is that the predictive distribution of the next observation is a mixture of predictive distributions that are defined by different components of the past and current data. This can be used to define a different parsimonious special case of the second-order Markov chain model that takes account of repeating patterns. The basic idea is that, if the past and current states belong to the pattern, the next state either continues the pattern with probability α or else is randomly generated according to a Markov chain of order less than that of the pattern, with probability $1 - \alpha$. If the previous states do not belong to the pattern, the next observation is generated randomly from the same Markov chain.

For the pattern 1312 with a Markov chain of order 0, this model is defined as follows. Let $\mathbf{i} = (i_0 \,|\, i_1, i_2)$. Then

$$p(\mathbf{i}) = \begin{cases} \alpha & \text{if } \mathbf{i} \in A, \\ (1-\alpha)\pi_{i_0}/\Sigma_{j:(j|i_1,i_2)\notin A}\pi_j & \text{if } \mathbf{i} \in B, \\ \pi_{i_0} & \text{otherwise,} \end{cases} \tag{12}$$

where $A = \{2|13, 1|21, 3|12, 1|31\}$, $B = \{\mathbf{i} \notin A : \exists \mathbf{j} \in A$ such that $j_1 = i_1,\ j_2 = i_2\}$ and $0 \leqslant \pi_j \leqslant 1$ $(j = 1, 2, 3)$, $\pi_1 + \pi_2 + \pi_3 = 1$.

Exact maximum likelihood estimators are available in analytical form for this model, namely

$$\hat{\alpha} = \frac{n_A}{n_A + n_B},$$

$$\hat{\pi}_j = \frac{1}{n - n_A} \sum_{\mathbf{i} \notin A : i_0 = j} n(\mathbf{i}).$$

The model is easily generalized to allow for a Markov chain of order 1 rather than order 0 in the non-pattern part of the model.

From Table 13, we see that neither of these models (models 7 and 8 in Table 13) fit as well as the full second-order Markov chain. The reason for this is that there is a second, less dominant, repeating pattern in the data, namely 112, which occurs 40 times. This is also a second-order pattern, defined by the transitions $2|11$, $1|21$, $1|12$. It is quite straightforward to model both transitions at once: the data are assumed to be generated by a mixture of the two patterns and a low order Markov chain. There are two minor complications: the transition $1|21$ is common to both patterns and the conditioning sequence 12 appears in both patterns, but with different succeeding phrases (3 for 1312 and 1 for 112). We deal with these by counting the transition $1|12$ as part of the 1312 pattern only, and by assigning separate parameters to the transition probabilities $p(j|12)$ $(j = 1, 2, 3)$.

The full model is therefore

$$p(\mathbf{i}) = \begin{cases} \alpha_h & \text{if } \mathbf{i} \in A_h, \\ (1-\alpha_h)\pi_{i_0}/\Sigma_{j:(j|i_1,i_2)\notin A_h}\pi_j & \text{if } \mathbf{i} \in B_h, \\ \gamma_{i_0} & \text{if } i_1 = 1,\ i_2 = 2, \\ \pi_{i_0} & \text{otherwise} \end{cases} \tag{13}$$

for $h = 1, 2$, where $A_1 = \{2|13,\ 1|21,\ 1|31\}$, $A_2 = \{2|11\}$ and $B_h = \{\mathbf{i} \notin A_h$: $\exists\, \mathbf{j} \in A_h$ with $j_1 = i_1, j_2 = i_2\}$. Here $h = 1$ corresponds to the 1312 pattern and $h = 2$ corresponds to the 112 pattern. The model is easily generalized to a Markov chain of order 1 for the non-pattern part, or of order -1 (the equiprobable case, obtained by constraining $\pi_1 = \pi_2 = \pi_3$). Exact maximum likelihood estimators are available in analytical form for each of these models.

Model (13) and its equiprobable specialization (models 9 and 10 in Table 13) have substantially better BIC values than the full second-order Markov chain model; the equiprobable model (model 9) is the best according to BIC. It is extremely parsimonious, using only four parameters to describe the full second-order structure. For this model, the parameter estimates are $\hat{\alpha}_1 = 0.98$, $\hat{\alpha}_2 = 0.78$, $\hat{\gamma}_1 = 0.16$, $\hat{\gamma}_2 = 0.08$ and $\hat{\gamma}_3 = 0.76$. Note in particular the value of $\hat{\alpha}_1$ close to 1, indicating the strong persistence of the 1312 pattern. The 112 pattern is also persistent, but much less so, as indicated by the lower $\hat{\alpha}_2$-value of 0.78.

The basic idea here may have potential for parsimoniously representing complex repeating patterns embedded in noise in applications such as speech recognition (where the number of states is large) and coding regions in DNA.

6. Discussion

Various generalizations of the MTD model have been proposed. Raftery (1985a, b) proposed ways of modelling the case where $m = \infty$, such as when the observations are counts. Adke and Deshmukh (1988) showed that asymptotic properties valid when m is finite also apply when $m = \infty$. It seems that our estimation method will work in that case also, provided that the (now doubly infinite) matrix Q is modelled parametrically. If $m = \infty$ and

$$\lim_{i \to \infty} \inf \{q(i|j)\} = 0 \quad \forall\, j, \tag{14}$$

then constraints (5) are equivalent to the positivity assumption (6), and the computational problem is greatly simplified.

Mehran (1989) considered the infinite lag MTD model, $l = \infty$, where λ_j is a parametric function of j. Our method seems applicable in this case also, although calculating the likelihood, or the fitted values for minimum χ^2 estimation, seems difficult. It may be possible to model discrete-valued time series with the long memory property by using this approach, by setting the λ_j equal to the π-weights for the fractionally differenced autoregressive integrated moving average ARIMA(p, d, q) process. Various continuous-valued environmental time series such as wind speeds are of this kind (Haslett and Raftery, 1989), and it seems reasonable to suppose that some discrete-valued time series might have this property also.

Martin and Raftery (1987) and Adke and Deshmukh (1988) pointed out that the MTD model remains well defined for arbitrary state spaces, which need not be finite, countable or even discrete. Equations (1) and (2) remain valid if p and q are interpreted as conditional densities, where q will usually have some parametric form. Le *et al.* (1990) have shown that this provides a framework for modelling bursts, outliers and flat stretches in continuous-valued time series, and that it is also good at modelling time series that are well fitted by conventional Gaussian autoregressive moving average models. If condition (14) holds, then so does positivity assumption (6). However, this is not always the case, even when the state space is continuous. For example, in continuous-valued directional time series, condition (14) does not necessarily hold. Craig (1989) has investigated MTD and other models for this situation, and has studied the consequences if condition (14) does not hold.

Our code, which is called MTD, was written in Fortran and may be obtained free of charge from Statlib by sending an electronic mail message to the Internet address statlib@lib.stat.cmu.edu consisting of the single line 'send mtd from general'. It calls the Numerical Algorithms Group (1991) subroutine E04UCF, which must be available to it.

Acknowledgements

Adrian E. Raftery was supported in part by Office of Naval Research contracts N-00014-88-K-0265 and N-00014-91-J-1074. Simon Tavaré was supported in part by

National Science Foundation grants DMS88-03284 and DMS90-05833 and National Institutes of Health grant GM 41746. The authors are grateful to Brian Francis for helpful discussions about GLIM, to Jia Ye for help with the computations, to Peter Alfeld for several discussions about hashing, to Jim Schimert, Jim Hughes and Chris Fraley for helpful discussions about optimization, to Ben Koop for sharing his data prior to publication and to the Editor and two referees for very helpful comments.

References

Adke, S. R. and Deshmukh, S. R. (1988) Limit distribution of a high order Markov chain. *J. R. Statist. Soc.* B, **50**, 105–108.

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM*, ch. 6. Oxford: Clarendon.

Avery, P. J. (1987) The analysis of intron data and their use in the detection of short signals. *J. Mol. Evoln*, **26**, 335–340.

Baker, R. J. (1986) *The GLIM Release 3.77 Reference Guide*. Oxford: Numerical Algorithms Group.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Am. Statist. Ass.*, **82**, 112–122.

Billingsley, P. (1961) *Statistical Inferences for Markov Processes*. Chicago: University of Chicago Press.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis*. Cambridge: Massachusetts Institute of Technology Press.

Blaisdell, B. E. (1983) A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nuclear DNA sequences. *J. Mol. Evoln*, **19**, 122–133.

Borodovsky, M. Y., Sprizhitsky, Y., Golovanov, E. and Alexandrov, A. (1986) Statistical patterns in the primary structures of functional regions in the genome of *E. Coli*: II, Nonuniform Markov models. *Mol. Biol.*, **20**, 1024–1033.

Chatfield, C. and Lemon, R. E. (1970) Analysing sequences of behavioral events. *J. Theoret. Biol.*, **29**, 427–445.

Craig, P. (1989) Time series analysis of directional data. *PhD Thesis*. Department of Statistics, Trinity College, Dublin.

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

Curnow, R. N. and Kirkwood, T. B. L. (1989) Statistical analysis of deoxyribonucleic acid sequence data—a review. *J. R. Statist. Soc.* A, **152**, 199–220.

Dawid, A. P. (1984) Statistical theory—the prequential approach. *J. R. Statist. Soc.* A, **147**, 278–292.

——— (1986) Probability forecasting. In *Encyclopaedia of Statistical Sciences* (eds S. Kotz and N. L. Johnson), vol. 7, pp. 210–218. New York: Wiley.

Doolittle, R. F. (ed.) (1990) Molecular evolution: computer analysis of protein and nucleic acid sequences. *Meth. Enzymol.*, **183**.

Good, I. J. (1952) Rational decisions. *J. R. Statist. Soc.* B, **14**, 107–114.

Haslett, J. and Raftery, A. E. (1989) Space-time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussion). *Appl. Statist.*, **38**, 1–50.

Jeffreys, H. (1961) *Theory of Probability*, 3rd edn. Oxford: Clarendon.

Karlin, S., Ost, F. and Blaisdell, B. E. (1988) Patterns in DNA and amino acid sequences and their statistical significance. In *Mathematical Methods for DNA Sequences* (ed. M. S. Waterman), pp. 133–157. Boca Raton: Chemical Rubber Company.

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

Kemeny, J. G. and Snell, J. L. (1960) *Finite Markov Chains*. New York: Van Nostrand.

Koop, B. F., Wilson, R. K., Wang, K., Vernooij, B., Zaller, D., Lam, C., Seto, D. and Hood, L. (1992) Organization, structure and function of 95kb of DNA spanning the murine T-cell receptor $C\alpha C/\delta$ region. Submitted to *Genomics*.

Kwok, M. C. O. (1988) Some results on higher order Markov chain models. *MPhil Thesis*. University of Hong Kong, Hong Kong.

Le, N. D., Martin, R. D. and Raftery, A. E. (1990) Modeling outliers, bursts and flat stretches in

times series using Mixture Transition Distribution (MTD) models. *Technical Report 194*. Department of Statistics, University of Washington, Seattle.

Li, W. K. and Kwok, M. C. O. (1989) Some results on the estimation of a higher order Markov chain.

Martin, R. D. and Raftery, A. E. (1987) Outliers, computation and non-Euclidean models. *J. Am. Statist. Ass.*, **82**, 1044–1050.

Mehran, F. (1989) Analysis of discrete longitudinal data: infinite lag Markov models. In *Statistical Data Analysis and Inference* (ed. Y. Dodge), pp. 533–541. New York: Elsevier.

Numerical Algorithms Group (1991) *Fortran Library, Mark 15*. Oxford: Numerical Algorithms Group.

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

——— (1985b) A new model for discrete-valued time series: autocorrelations and extensions. *Rass. Met. Statist. Appl.*, **3–4**, 149–162.

——— (1986a) A note on Bayes factors for log-linear contingency table models with vague prior information. *J. R. Statist. Soc.* B, **48**, 249–250.

——— (1986b) Choosing models for cross-classifications. *Am. Sociol. Rev.*, **51**, 145–146.

——— (1988) Approximate Bayes factors for generalized linear models. *Technical Report 121*. Department of Statistics, University of Washington, Seattle.

Raftery, A. E., Haslett, J. and McColl, E. (1982) Wind power: a space–time process? In *Time Series Analysis: Theory and Practice* (ed. O. D. Anderson), pp. 191–202. Amsterdam: North-Holland.

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

Tavaré, S. and Giddings, B. W. (1988) Some statistical aspects of the primary structure of nucleotide sequences. In *Mathematical Methods for DNA Sequences* (ed. M. S. Waterman), pp. 117–132. Boca Raton: Chemical Rubber Company.

Thomas, M. U. and Barr, D. R. (1977) An approximate test for Markov chain lumpability. *J. Am. Statist. Ass.*, **72**, 175–179.

Waterman, M. S. (ed.) (1988) *Mathematical Methods for DNA Sequences*. Boca Raton: Chemical Rubber Company.

Watterson, G. A. (1992) A stochastic analysis of three viral sequences. *Mol. Biol. Evoln*, **9**, 666–677.

Wilson, R. K., Koop, B. F., Chen, C., Halloran, N., Sclammis, R. and Hood, L. (1992) Nucleotide sequence analysis of 95kb near the 3′ end of the murine T-cell receptor α/δ chain locus: strategy and methodology. Submitted to *Genomics*.