Statistical Science
2003, Vol. 18, No. 3, 357–376

Regression Theory for Categorical Time Series

Konstantinos Fokianos and Benjamin Kedem

Abstract. Categorical—or qualitative—time series data with random time-dependent covariates are frequently encountered in diverse applications as the list of examples shows. As with "ordinary" time series, the data analyst is faced with the same problems of modeling, estimation, model checking, diagnostics and prediction. The present work shows that these questions can be attacked by means of regression theory for categorical time series whose foundation is based on generalized linear models and partial likelihood inference. A variety of models are provided to illustrate the selection of the link function and recent large sample results are reviewed. The theory is developed without resorting to the Markov assumption and to the notion of stationarity. Moreover, regression methods for categorical time series allow for parsimonious modeling and incorporation of random time-dependent covariates as opposed to other procedures. In particular, nominal and ordinal time series are analyzed and compared empirically to Markov chains and mixture transition distribution models.

Key words and phrases: Random time-dependent covariates, partial likelihood, martingale, multinomial logits, proportional odds, link function, deviance, residuals, Markov chain, mixture transition distribution model.

1. INTRODUCTION

Figure 1 displays the first 300 records of EEG sleep state scores, typical of newborn infants, classified or quantized in four categories as follows:

(1) quiet sleep,
(2) indeterminate sleep,
(3) active sleep,
(4) awake.

Here the sleep state categories or levels are assigned integer values. This is an example of a *categorical time series* $\{Y_t\}$, $t = 1, \ldots, N$, taking the values $1, \ldots, 4$. These data are accompanied by *random time-dependent covariates*, that is, measurements of heart rate and temperature—see Section 6.3. The plot raises several questions. Is there an apparent "periodic" tendency in the data? What is the best way to predict a future sleep state? Do lagged values of sleep state determine future states? Can the covariates be used to predict sleep states? These questions and others show that "ordinary" and categorical time series pose the same basic problems. The present article offers some answers to these problems by considering *regression models for categorical time series*.

The long list of references at the end of the article shows that a number of different strategies have been proposed for modeling of categorical time series over the last 20 years or so. Different models include Markov chain models, integer autoregressive processes, discrete ARMA models and so on. However, a successful and versatile approach to the problem of regression modeling for categorical time series utilizes the simple and elegant theory of generalized linear models. Accordingly, it is sufficient to express the conditional expectation of the response as a function of autoregressive components, past observations and, more generally, random time-dependent

Konstantinos Fokianos is Assistant Professor, Department of Mathematics and Statistics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus (e-mail: fokianos@ucy.ac.cy). Benjamin Kedem is Professor, Department of Mathematics, University of Maryland, College Park, Maryland 20742-4015 (e-mail: bnk@math.umd.edu).

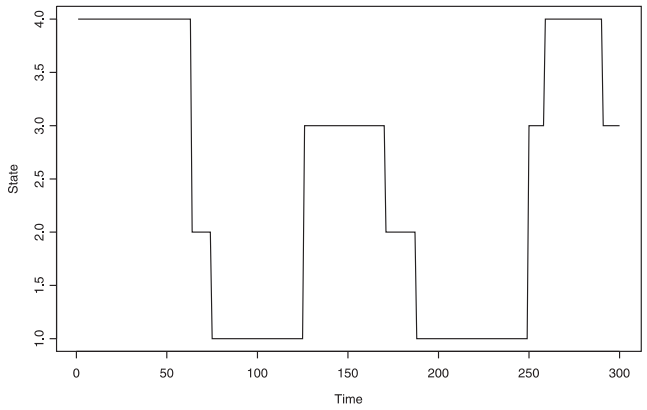

FIG. 1. *First* 300 *observations of the sleep data.*

covariates. *A gained advantage is that neither the Markov property nor stationarity is assumed*. Furthermore, experience shows that both positive and negative association can be taken into account by a suitable parametrization of the model. Sections 2 and 3 take up the issues of modeling and link selection for regression models of categorical time series. In particular, a detailed exposition of models for nominal and ordinal time series is offered.

Estimation theory appeals to the methodology of partial likelihood—an important inferential method for dependent data. As it turns out, partial likelihood allows for sequential inference with respect to a filtration generated by all the information available to the data analyst at the time of observation. Estimation, diagnostics, forecasting and model checking are carried out easily where the computations are implemented by a number of existing software packages. The problem of estimation and testing is discussed in Section 4 where recent progress in this area is reported—see Theorems 4.1 and 4.2—showing that standard results of likelihood analysis are carried over to the dependent data case.

The reader should recognize that Markov chains provide a simple but important example of categorical time series where lagged values of the response are instrumental in determining its future states. The topic of Markov chains has been studied by many authors (see, e.g., Karlin and Taylor, 1975; Meyn and Tweedie, 1993). The texts by Billingsley (1961), Basawa and Prakasa Rao (1980, Chapter 4) and Guttorp (1995, Chapter 2) present statistical inference theory for Markov chains.

Recall that a process $\{Y_t\}$, $t = 1, \ldots, N$, defined on $\{1, 2, \ldots, m\}$, is called a Markov chain of order p if it satisfies

$$\begin{aligned} &\mathrm{P}[Y_t = k | Y_{t-1}, Y_{t-2}, \ldots] \\ &\quad = \mathrm{P}[Y_t = k | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}], \\ &\qquad\qquad k = 1, \ldots, m. \end{aligned} \tag{1}$$

Thus, given the past values of $Y_{t-1}, \ldots, Y_{t-p}$, (1) provides the conditional probabilities of a future state or category k. Markov modeling in the context of categorical time series can be problematic for two reasons. First, as the order of the Markov chain increases so does the number of free parameters. Indeed, for a Markov chain of order p, the total number of parameters needed to be estimated is equal to $m^p(m-1)$. Second, the insistence on using the Markov property requires the specification of the joint dynamics of the response and the covariates. This, however, may not always be possible.

An alternative approach, among others, to Markov chain models has been the *mixture transition distribution* model—a topic presented in Section 5.1. Although the problem of exponentially increasing parameters is cleverly bypassed, the issue of incorporating covariates still remains unresolved for the mixture transition distribution model.

The examples of Section 6 show the flexibility of regression theory for categorical time series especially when compared to the Markov models. The first example in Section 6.1 refers to explanatory analysis of DNA sequence data. Viewing the sequence of DNA letters as a nominal time series, the regression methodology can discover dependencies in the data which cannot be assessed by a Markov model. The next example in Section 6.2 shows how the regression theory is employed in discovering independence, while the last example in Section 6.3—discussed briefly above—shows how random time-dependent covariates can be taken into account. The presentation concludes with some other topics related to categorical time series.

2. MODELING

Assume that we observe a categorical time series $\{Y_t\}$, $t = 1, \ldots, N$, and let m be the number of categories. In other words, for each t, the possible values of Y_t are $1, 2, \ldots, m-1, m$, where the "first" category is assigned the integer value of 1, the "second" category is assigned the integer value of 2 and so on. In general, the assignment of integer values to the categories is a matter of convenience and hence it is not unique.

To reduce the amount of arbitrariness incurred by integer assignment to categories, it is helpful to note that the tth observation of any categorical time series—regardless of the measurement scale—can be expressed by the vector $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{tq})'$ of length $q = m - 1$, with elements

$$Y_{tj} = \begin{cases} 1, & \text{if the } j\text{th category} \\ & \text{is observed at time } t, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

for $t = 1, \ldots, N$ and $j = 1, \ldots, q$. Denote by $\boldsymbol{\pi}_t = (\pi_{t1}, \ldots, \pi_{tq})'$ the vector of conditional probabilities given $\mathcal{F}_{t-1}$, where

$$\pi_{tj} = \mathrm{E}[Y_{tj}|\mathcal{F}_{t-1}] = \mathrm{P}(Y_{tj} = 1|\mathcal{F}_{t-1}), \quad j = 1, \ldots, q,$$

for every $t = 1, \ldots, N$. At times, we refer to the π_{tj} as "transition probabilities." Define

$$Y_{tm} = 1 - \sum_{j=1}^{q} Y_{tj}$$

and

$$\pi_{tm} = 1 - \sum_{j=1}^{q} \pi_{tj}.$$

The σ-field $\mathcal{F}_{t-1}$ is generated by $\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots, \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \ldots$, that is,

$$\mathcal{F}_{t-1} = \sigma\{\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots, \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \ldots\},$$

where $\{\mathbf{Z}_{t-1}\}$, $t = 1, \ldots, N$, stands for a $p \times q$ *matrix* that represents a covariate process. In other words, each response Y_{tj} corresponds to a vector of length p of random time-dependent covariates which forms the jth column of $\mathbf{Z}_{t-1}$. The covariate matrix may consist of lagged values of the response process and of any other auxiliary process *known to the observer at time* t.

Following the theory of generalized linear models, McCullagh and Nelder (1989) assume that the vector of transition probabilities—that is, the conditional expectation of the response vector—is linked to the covariate process through the equation

$$\boldsymbol{\pi}_t(\boldsymbol{\beta}) = \mathbf{h}(\mathbf{Z}'_{t-1}\boldsymbol{\beta}), \tag{3}$$

with $\boldsymbol{\beta}$ a p-dimensional vector of time-invariant parameters. Equation (3) gives the general form of a *multivariate* generalized linear model for categorical time series

$$\boldsymbol{\pi}_t(\boldsymbol{\beta}) = \begin{pmatrix} \pi_{t1}(\boldsymbol{\beta}) \\ \pi_{t2}(\boldsymbol{\beta}) \\ \vdots \\ \pi_{tq}(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} h_1(\mathbf{Z}'_{t-1}\boldsymbol{\beta}) \\ h_2(\mathbf{Z}'_{t-1}\boldsymbol{\beta}) \\ \vdots \\ h_q(\mathbf{Z}'_{t-1}\boldsymbol{\beta}) \end{pmatrix} = \mathbf{h}(\mathbf{Z}'_{t-1}\boldsymbol{\beta}),$$

where the *inverse* link function $\mathbf{h}$ is defined on $\mathbb{R}^q$ and takes values in $\mathbb{R}^q$ as well. To guarantee that the transition probabilities fall between 0 and 1, we impose the condition that $\mathbf{h}$ maps a subset $H \subseteq \mathbb{R}^q$ one to one onto $\{(w_1, \ldots, w_q)' : w_j > 0, j = 1, \ldots, q, \sum_{j=1}^{q} w_j < 1\}$. Model (3) has been considered by a number of authors, including Fahrmeir and Kaufmann (1987), Kaufmann (1987), Pruscha (1993) and Fokianos and Kedem (1998), where the past probability vector $\boldsymbol{\pi}_{t-1}$ is included as a covariate. Some further work can be found in Brillinger (1996) and Fokianos, Kedem and Short (1996). In addition, recent work by Brillinger, Morettin, Irizarry and Chiann (2000) develops a wavelet-based method for the analysis of categorical time series.

3. LINK FUNCTIONS FOR CATEGORICAL TIME SERIES

We now turn to the problem of what constitutes a reasonable choice for the inverse link function $\mathbf{h}$ in the context of regression models for categorical time series. We introduce some widely used models for the analysis of categorical time series, including the so-called *multinomial logit* and *cumulative odds* models. In general, the choice of model depends on one of three measurement scales: *nominal*, *ordinal* and *interval*. We shall only examine nominal and ordinal time series since interval (e.g., quantized) time series can be handled by methods designed for ordinal data.

3.1 Models for Nominal Time Series

By nominal categorical variables, we mean variables whose scale of measurement lacks any natural ordering. For instance, the daily choice of transportation is an example of a nominal time series. The multinomial logit model defined by Agresti (1990, Section 9.2),

$$\pi_{tj}(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}'_j \mathbf{z}_{t-1})}{1 + \sum_{l=1}^{q} \exp(\boldsymbol{\beta}'_l \mathbf{z}_{t-1})}, \quad j = 1, \ldots, q, \tag{4}$$

is frequently employed in the analysis of nominal time series. Here $\boldsymbol{\beta}_j$, $j = 1, \ldots, q$, are d-dimensional regression parameters and $\mathbf{z}_{t-1}$ is a corresponding d-dimensional vector of stochastic time-dependent covariates independent of j. Obviously,

$$\pi_{tm}(\boldsymbol{\beta}) = \frac{1}{1 + \sum_{l=1}^{q} \exp(\boldsymbol{\beta}'_l \mathbf{z}_{t-1})}.$$

Typical examples of $\mathbf{z}_{t-1}$ include

$$\mathbf{z}_{t-1} = \big(1, W_t, Y_{t-1}, \log(W_{t-1})\big)'$$

or, when interactions are entertained,

$$\mathbf{z}_{t-1} = (1, W_{t-1}, Y_{t-1}, Y_{t-1}W_{t-1})'$$

and so on given an auxiliary process $\{W_t\}$.

The multinomial logit model (4) is derived either by a straightforward extension of the logistic model or by maximizing a random utility. The first approach defines log-odds ratios relative to π_{tm}:

$$\log \frac{\pi_{tj}}{\pi_{tm}} = \boldsymbol{\beta}_j' \mathbf{z}_{t-1}, \quad j = 1, \ldots, q.$$

Then (4) follows from the fact that $\sum_{j=1}^{m} \pi_{tj} = 1$. The second line of argument uses the maximization of a random utility function along the lines of McFadden (1973).

An essential observation is that (4) implies

$$\log \frac{\pi_{tj}}{\pi_{ti}} = (\boldsymbol{\beta}_j' - \boldsymbol{\beta}_i')\mathbf{z}_{t-1}.$$

Thus, the ratio π_{tj}/π_{ti} for the jth and ith categories is the same regardless of the total number of categories m. This property is referred to as *independence of irrelevant alternatives* (Luce, 1959).

Model (4) is a special case of (3). Indeed, define β to be the qd vector

$$\beta = (\beta_1', \ldots, \beta_q')',$$

and $\mathbf{Z}_{t-1}$ the $qd \times q$ matrix

$$\mathbf{Z}_{t-1} = \begin{bmatrix} \mathbf{z}_{t-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{t-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{z}_{t-1} \end{bmatrix}.$$

Let $\mathbf{h}$ stand for the vector-valued function whose components h_j, $j = 1, \ldots, q$, are given by

$$\pi_{tj}(\boldsymbol{\beta}) = h_j(\boldsymbol{\eta}_t) = \frac{\exp(\eta_{tj})}{1 + \sum_{l=1}^{q} \exp(\eta_{tl})}, \quad j = 1, \ldots, q,$$

with

$$\boldsymbol{\eta}_t = (\eta_{t1}, \ldots, \eta_{tq})' = \mathbf{Z}_{t-1}'\boldsymbol{\beta}.$$

With this notation, (3) reduces to (4) when $p = qd$.

3.1.1 *Example*: *multinomial logit model with a periodic component.* It is instructive to examine closely a simple example by means of simulated data. Figure 2(a) displays a typical realization of a categorical time series with $m = 3$ categories and length $N = 200$. Since $m = 3$, $\mathbf{Y}_t$ has $q = 2$ components: $\mathbf{Y}_t = (Y_{t1}, Y_{t2})'$. The data have been generated accord-

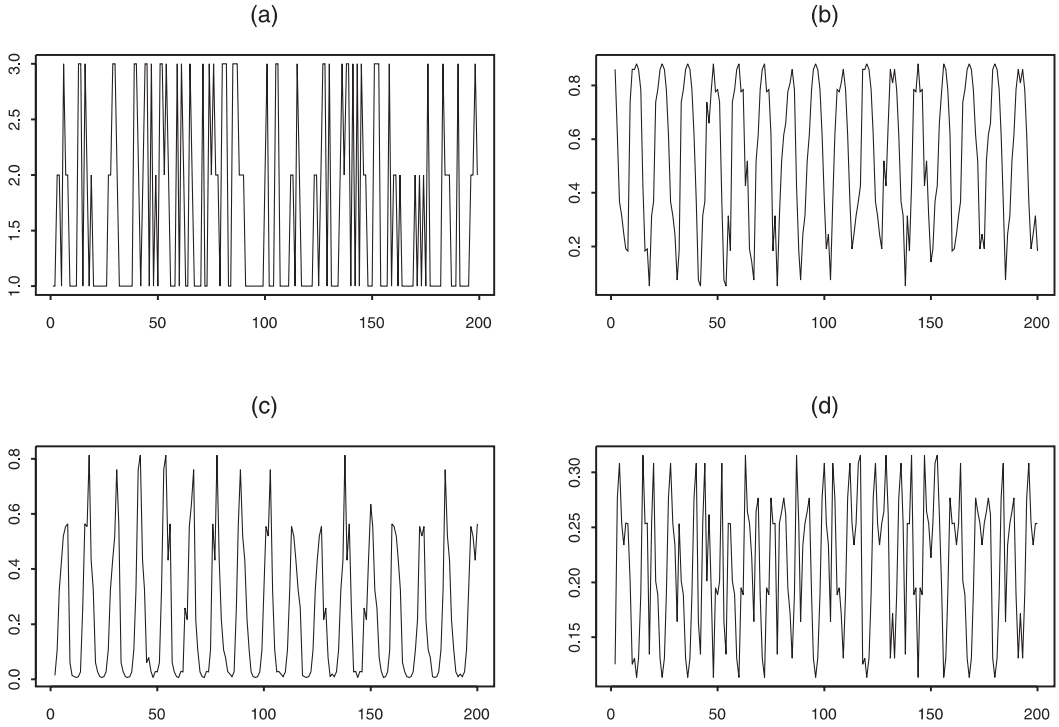

FIG. 2. *Typical realization of the multinomial logit model* (4) *with three categories. Here* $\boldsymbol{\beta}_1 = (0.30, 1.25, 0.50, 1.00)'$, $\boldsymbol{\beta}_2 = (-0.20, -2.00, -0.75, -1.00)'$, $\mathbf{z}_{t-1} = (1, \cos(2\pi t/12), \mathbf{Y}_{t-1}')'$ *and* $N = 200$. (a) Y_t, (b) π_{t1}, (c) π_{t2}, (d) π_{t3}.

ing to the model

$$\begin{aligned}\log\left(\frac{\pi_{t1}}{\pi_{t3}}\right) &= \boldsymbol{\beta}_1'\mathbf{z}_{t-1}\\ &= \beta_{10} + \beta_{11}\cos\left(\frac{2\pi t}{12}\right)\\ &\quad + \beta_{12}Y_{(t-1)1} + \beta_{13}Y_{(t-1)2},\end{aligned}$$

(5)

$$\begin{aligned}\log\left(\frac{\pi_{t2}}{\pi_{t3}}\right) &= \boldsymbol{\beta}_2'\mathbf{z}_{t-1}\\ &= \beta_{20} + \beta_{21}\cos\left(\frac{2\pi t}{12}\right)\\ &\quad + \beta_{22}Y_{(t-1)1} + \beta_{23}Y_{(t-1)2},\end{aligned}$$

with $\boldsymbol{\beta}_1 = (0.30, 1.25, 0.50, 1.00)'$, $\boldsymbol{\beta}_2 = (-0.20, -2.00, -0.75, -1.00)'$ and $\mathbf{z}_{t-1} = (1, \cos(2\pi t/12), \mathbf{Y}_{t-1}')'$. In other words, the simulated model incorporates a sinusoidal component and a lagged value of order 1. One starts with arbitrary values for the Y_{0j} to get the π_{1j} and then uses the π_{1j} to generate the Y_{1j} and so on. Figure 2(b)–(d) displays the transition probabilities of each of the categories, respectively. Figure 3 displays the sample autocorrelation function of the simulated data. The upper left and lower right panels display plots of the sample autocorrelation functions of Y_{t1} and Y_{t2}, respectively. The other plots depict the sample cross-correlation function between Y_{t1} and Y_{t2} for positive (upper right) and negative (lower left) lags. In all these plots, the sinusoidal component is apparent. Notice that in Figure 2(a) the assignment of values to the three categories, namely 1, 2, 3, is arbitrary, but this has no bearing on the final results due to (2).

3.2 Models for Ordinal Time Series

Ordinal categorical variables—such as blood pressure classified as low, normal and high—are measured on a scale endowed with a natural ordering. Thus, the hourly blood pressure of an individual charted as low, normal and high constitutes an ordinal time series. The cumulative odds model (Snell, 1964; McCullagh, 1980) is often used in applications for the analysis of ordinal data. The derivation of this model is better understood by means of a latent or auxiliary variable. That is, we assume the observed data result from the following threshold mechanism. Let

$$X_t = -\boldsymbol{\gamma}'\mathbf{z}_{t-1} + e_t,$$

where e_t is a sequence of i.i.d. random variables with continuous c.d.f. F, $\boldsymbol{\gamma}$ is a d-dimensional vector of parameters and $\mathbf{z}_{t-1}$ is a covariate vector of the same dimension. The process $\{X_t\}$, referred to as a "latent" process, may or may not be observed, but regardless of whether it is observed or not, the same calculations persist. Define a categorical time series

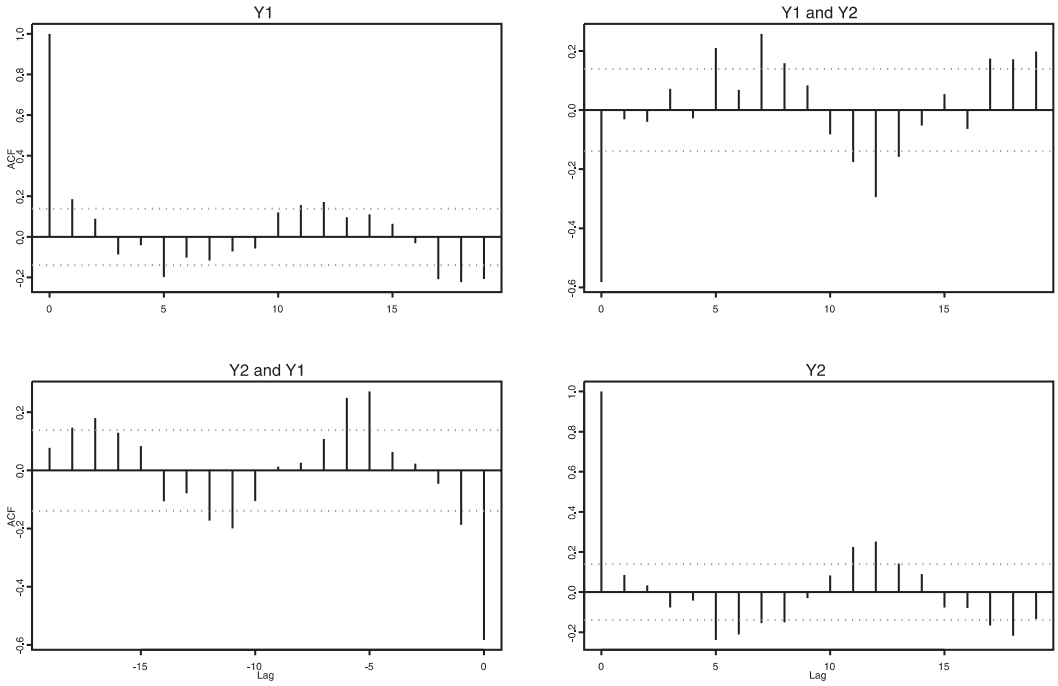

FIG. 3. *Sample autocorrelation and cross-correlation functions of the simulated data from Figure* 2.

$\{Y_t\}$, $t = 1, \ldots, N$, from the levels of $\{X_t\}$,

$$Y_t = j \quad \Longleftrightarrow \quad Y_{tj} = 1 \quad \Longleftrightarrow \quad \theta_{j-1} \le X_t < \theta_j$$

for $j = 1, \ldots, m$, where $\{\theta_0, \theta_1, \ldots, \theta_m\}$ is a set of *threshold parameters* satisfying

$$-\infty = \theta_0 < \theta_1 < \cdots < \theta_m = \infty.$$

Then

$$\begin{aligned} \pi_{tj} &= \mathrm{P}(\theta_{j-1} \le X_t < \theta_j | \mathcal{F}_{t-1}) \\ &= F(\theta_j + \boldsymbol{\gamma}' \mathbf{z}_{t-1}) - F(\theta_{j-1} + \boldsymbol{\gamma}' \mathbf{z}_{t-1}) \end{aligned} \tag{6}$$

for $j = 1, \ldots, m$. In other words,

$$\begin{aligned} &\mathrm{P}(Y_t \le j | \mathcal{F}_{t-1}) \\ &\quad = F(\theta_j + \boldsymbol{\gamma}' \mathbf{z}_{t-1}), \quad j = 1, \ldots, m. \end{aligned} \tag{7}$$

From the estimates of (6), we obtain estimates for (7), since the set of the cumulative probabilities corresponds one to one to the set of response probabilities. Many different special cases arise for various choices for F. For example, the cumulative logistic or *proportional odds* model is obtained when F is the logistic distribution function,

$$F_l(x) = \frac{1}{1 + \exp(-x)}.$$

Then we have

$$\log\left\{\frac{\mathrm{P}[Y_t \le j | \mathcal{F}_{t-1}]}{\mathrm{P}[Y_t > j | \mathcal{F}_{t-1}]}\right\} = \theta_j + \boldsymbol{\gamma}' \mathbf{z}_{t-1} \tag{8}$$

for $j = 1, \ldots, q$. Other choices for F include the standard normal cumulative distribution function

$$F \equiv \Phi,$$

the extreme minimal distribution function

$$F \equiv 1 - \exp\bigl(-\exp(x)\bigr)$$

and the extreme maximal distribution function

$$F \equiv \exp\bigl(-\exp(-x)\bigr).$$

In principle, any inverse link function appropriate for binary data can be used when entertaining a cumulative odds model.

To recognize that model (7) is a special case of (3), let $\boldsymbol{\beta}$ denote the $q + d$ vector

$$\boldsymbol{\beta} = (\theta_1, \ldots, \theta_q, \boldsymbol{\gamma}')'$$

and $\mathbf{Z}_{t-1}$ the $(q + d) \times q$ matrix

$$\mathbf{Z}_{t-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \mathbf{z}_{t-1} & \mathbf{z}_{t-1} & \cdots & \mathbf{z}_{t-1} \end{bmatrix}.$$

Now set

$$\mathbf{h} = (h_1, \ldots, h_q)',$$

with

$$\begin{aligned} \pi_{t1}(\boldsymbol{\beta}) &= h_1(\boldsymbol{\eta}_t) = F(\eta_{t1}), \\ \pi_{tj}(\boldsymbol{\beta}) &= h_j(\boldsymbol{\eta}_t) = F(\eta_{tj}) - F(\eta_{t(j-1)}), \\ &\qquad\qquad j = 2, \ldots, q, \end{aligned}$$

where

$$\boldsymbol{\eta}_t = (\eta_{t1}, \ldots, \eta_{tq})' = \mathbf{Z}_{t-1}' \boldsymbol{\beta}.$$

It is clear that (3) is satisfied with this notation and $p = q + d$.

Some other models worth mentioning for the analysis of ordinal responses include the *continuation ratio* model specified by

$$F^{-1}\left(\frac{\pi_{tj}(\boldsymbol{\beta})}{\pi_{t(j+1)}(\boldsymbol{\beta}) + \cdots + \pi_{tm}(\boldsymbol{\beta})}\right) = \boldsymbol{\beta}' \mathbf{z}_{t-1}, \tag{9}$$

and the *adjacent categories logit* model given by

$$\mathrm{P}(Y_t = j | Y_t \in \{r, r+1\}, \mathcal{F}_{t-1}) = F(\boldsymbol{\beta}' \mathbf{z}_{\mathbf{t-1}}), \tag{10}$$

where F stands for a continuous c.d.f. Various authors have considered the so-called *two-step* and *mean response* models—the latter for the analysis of interval response variables. The reader is referred to Agresti (1990), Johnson and Albert (1999) and Fahrmeir and Tutz (2001, Chapter 3) for further details on modeling aspects of ordinal and interval data.

3.2.1 *Example*: *proportional odds model with a periodic component.* Figure 4(a) shows a typical realization of an ordinal categorical time series of length $N = 200$ with $m = 3$ categories generated by the following proportional odds model (8):

$$\begin{aligned} &\log\left\{\frac{\mathrm{P}[Y_t \le 1 | \mathcal{F}_{t-1}]}{\mathrm{P}[Y_t > 1 | \mathcal{F}_{t-1}]}\right\} \\ &\quad = \theta_1 + \boldsymbol{\gamma}' \mathbf{z}_{t-1} \\ &\quad = \theta_1 + \gamma_1 \cos\left(\frac{2\pi t}{12}\right) + \gamma_2 Y_{(t-1)1} + \gamma_3 Y_{(t-1)2} \end{aligned}$$

and

$$\begin{aligned} &\log\left\{\frac{\mathrm{P}[Y_t \le 2 | \mathcal{F}_{t-1}]}{\mathrm{P}[Y_t > 2 | \mathcal{F}_{t-1}]}\right\} \\ &\quad = \theta_2 + \boldsymbol{\gamma}' \mathbf{z}_{t-1} \\ &\quad = \theta_2 + \gamma_1 \cos\left(\frac{2\pi t}{12}\right) + \gamma_2 Y_{(t-1)1} + \gamma_3 Y_{(t-1)2}. \end{aligned}$$

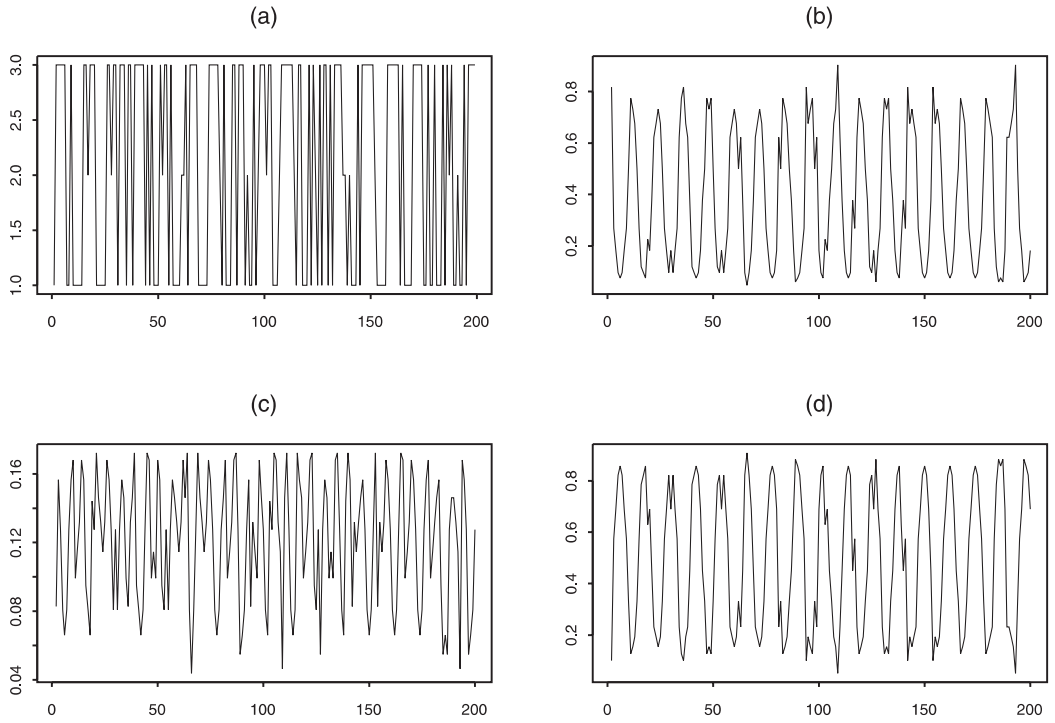

FIG. 4. *Typical realization of the proportional odds model* (8) *with three categories. Here* $\theta_1 = -0.50$, $\theta_2 = 0.20$, $\boldsymbol{\gamma} = (2.00, -0.50, 1.00)'$, $\mathbf{z}_{t-1} = (\cos(2\pi t/12), \mathbf{Y}'_{t-1})'$ *and* $N = 200$. (a) Y_t, (b) π_{t1}, (c) π_{t2}, (d) π_{t3}.

The model parameters are $\theta_1 = -0.50$, $\theta_2 = 0.20$, $\gamma = (2.00, -0.50, 1)'$, and the covariate vector $\mathbf{z}_{t-1} = (\cos(2\pi t/12), \mathbf{Y}'_{t-1})'$ consists of a sinusoidal component and a lagged value of order 1. Figure 4(b)–(d) displays the corresponding transition probabilities of each of the categories, respectively. Figure 5 displays the

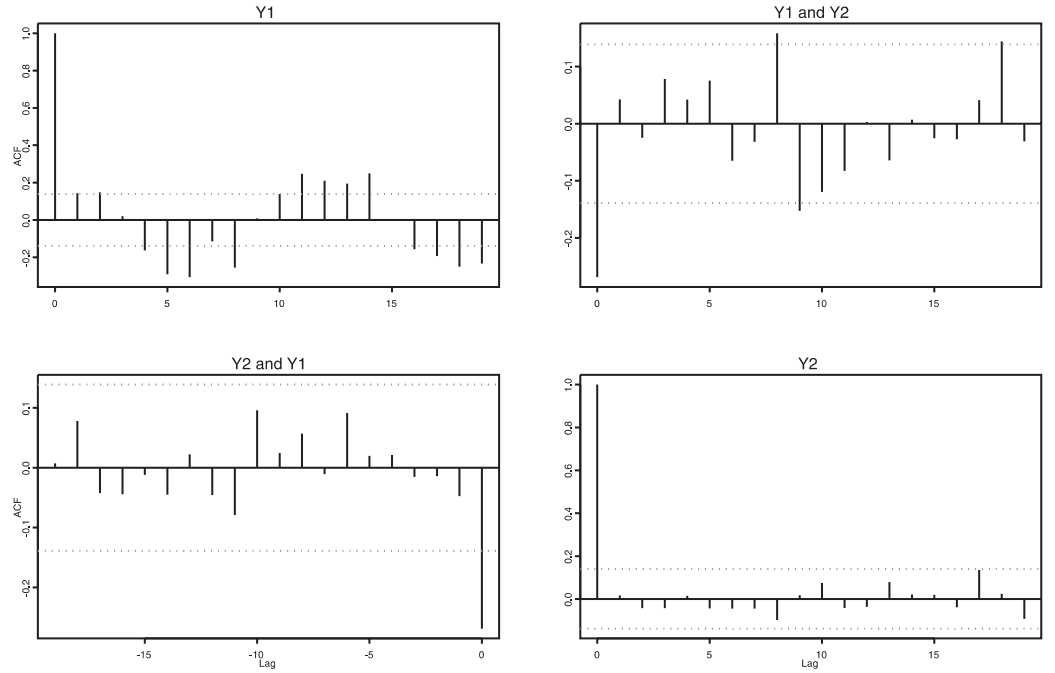

FIG. 5. *Sample autocorrelation and cross-correlation functions of the simulated data from Figure* 4.

sample autocorrelation function of the simulated data. The sinusoidal component is manifested clearly, especially in the upper left plot of Figure 5.

4. PARTIAL LIKELIHOOD ESTIMATION

The central statistical issue for a categorical time series regression model is to estimate the vector of parameters $\boldsymbol{\beta}$. Since the data are dependent, we attack the problem through the partial likelihood methodology, which was suggested by Cox (1975). Partial likelihood successfully approaches the problem of estimation and testing by means of martingale theory. It has been proved a useful tool for time series following generalized linear models (see, e.g., Wong, 1986; Slud and Kedem, 1994; Fokianos and Kedem, 1998, among others). According to Fahrmeir and Kaufmann (1987), Kaufmann (1987) and Fokianos and Kedem (1998), the partial likelihood (PL) function relative to $\boldsymbol{\beta}$, $\mathcal{F}_t$ and the data is given by

$$\begin{aligned}\mathrm{PL}(\boldsymbol{\beta}) &= \prod_{t=1}^{N} f(\mathbf{y}_t;\boldsymbol{\beta}|\mathcal{F}_{t-1}) \\ &= \prod_{t=1}^{N}\prod_{j=1}^{m} \pi_{tj}(\boldsymbol{\beta})^{y_{tj}},\end{aligned} \tag{11}$$

so that the partial log-likelihood is given by

$$l(\boldsymbol{\beta}) \equiv \log \mathrm{PL}(\boldsymbol{\beta}) = \sum_{t=1}^{N}\sum_{j=1}^{m} y_{tj}\log \pi_{tj}(\boldsymbol{\beta}). \tag{12}$$

It is useful to introduce the *logit* function at this point:

$$\begin{aligned}\mathbf{logit}(\mathbf{x}) = \Bigg(&\log\Bigg(\frac{x_1}{1-\sum_{j=1}^{q} x_j}\Bigg), \\ &\ldots, \log\Bigg(\frac{x_q}{1-\sum_{j=1}^{q} x_j}\Bigg)\Bigg),\end{aligned} \tag{13}$$

for a q-dimensional vector $\mathbf{x}$ which belongs in the set $\{(x_1,\ldots,x_q)' : x_j > 0, j = 1,\ldots,q, \sum_{j=1}^{q} x_j < 1\}$.

Computation of the maximum partial likelihood estimator (MPLE) $\hat{\boldsymbol{\beta}}$ is carried out by maximizing the partial log-likelihood (12). This, in turn, implies that if the MPLE $\hat{\boldsymbol{\beta}}$ exists then it is given as the solution of the partial score equations

$$\nabla l(\boldsymbol{\beta}) = \nabla \log \mathrm{PL}(\boldsymbol{\beta}) = \mathbf{0}, \tag{14}$$

assuming differentiability. The solution of the partial score equations (14) is obtained by Fisher scoring.

We obtain the *partial score* by differentiating (12)

$$\begin{aligned}\mathbf{S}_N(\boldsymbol{\beta}) &= \nabla l(\boldsymbol{\beta}) \\ &= \Bigg(\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1},\ldots,\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_p}\Bigg)' \\ &= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta})\big(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})\big).\end{aligned} \tag{15}$$

Set

$$\mathbf{U}_t(\boldsymbol{\beta}) = \mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}),$$

where

$$\mathbf{D}_t(\boldsymbol{\beta}) = \Bigg[\frac{\partial \mathbf{h}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t'}\Bigg]$$

and $\boldsymbol{\Sigma}_t(\boldsymbol{\beta})$ is the conditional covariance matrix of $\mathbf{Y}_t$ with generic elements

$$\sigma_t^{(ij)}(\boldsymbol{\beta}) = \begin{cases} -\pi_{ti}(\boldsymbol{\beta})\pi_{tj}(\boldsymbol{\beta}), & \text{if } i \neq j, \\ \pi_{ti}(\boldsymbol{\beta})\big(1-\pi_{ti}(\boldsymbol{\beta})\big), & \text{if } i = j, \end{cases}$$

for $i, j = 1,\ldots,q$. It follows that the partial score (15) can also be expressed by the equation

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})\big(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})\big), \tag{16}$$

where

$$\mathbf{U}_t(\boldsymbol{\beta}) = \Bigg[\frac{\partial \mathbf{u}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t'}\Bigg]$$

is now a $q \times q$ matrix and $\boldsymbol{\eta}_t = \mathbf{Z}_{t-1}'\boldsymbol{\beta}$. The function $\mathbf{u} = (u_1,\ldots,u_q)$ is the composition of the functions $\mathbf{h}$ in (3) and **logit** in (13). In other words,

$$\mathbf{u} = (u_1,\ldots,u_q) = \big(h_1(\mathbf{logit}),\ldots,h_q(\mathbf{logit})\big).$$

In what follows, we shall use (16) rather than (15).

The conditional information matrix is given by

$$\begin{aligned}\mathbf{G}_N(\boldsymbol{\beta}) &= \sum_{t=1}^{N} \mathrm{Cov}\big[\mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})\big(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})\big)|\mathcal{F}_{t-1}\big] \\ &= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t(\boldsymbol{\beta})\mathbf{U}_t'(\boldsymbol{\beta})\mathbf{Z}_{t-1}'.\end{aligned} \tag{17}$$

The unconditional information matrix is given by

$$\mathbf{F}_N(\boldsymbol{\beta}) = \mathrm{E}[\mathbf{G}_N(\boldsymbol{\beta})] \tag{18}$$

and the second derivative of the partial log-likelihood multiplied by -1 is

$$\mathbf{H}_N(\boldsymbol{\beta}) = -\nabla\nabla' l(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) - \mathbf{R}_N(\boldsymbol{\beta}), \tag{19}$$

where

$$\mathbf{R}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N}\sum_{r=1}^{q} \mathbf{Z}_{t-1}\mathbf{W}_{tr}(\boldsymbol{\beta})\mathbf{Z}'_{t-1}\big(Y_{tr} - \pi_{tr}(\boldsymbol{\beta})\big),$$

with

$$\mathbf{W}_{tr}(\boldsymbol{\beta}) = \left[\frac{\partial^2 u_r(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t\, \partial \boldsymbol{\eta}'_t}\right]$$

for $r = 1, \ldots, q$.

4.1 Large-Sample Theory

Asymptotic properties of the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ are examined via the score function and the conditional information matrix. It turns out that the following theorem holds under some mild regularity conditions; see Fokianos and Kedem (1998).

THEOREM 4.1. *Consider model* (3). *Then*, *under some mild regularity conditions* (*Assumption* A *of Fokianos and Kedem*, 1998), *we obtain the following*:

1. *There exists a locally unique maximum partial likelihood estimator*, $\hat{\boldsymbol{\beta}}$, *with probability tending to* 1 *as* $N \to \infty$.
2. *The estimator is consistent and asymptotically normal*,

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$$

and

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}\big(0, \mathbf{G}^{-1}(\boldsymbol{\beta})\big)$$

as $N \to \infty$.
3. *The following is true*:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{1}{\sqrt{N}}\mathbf{G}(\boldsymbol{\beta})^{-1} S_N(\boldsymbol{\beta}) \xrightarrow{p} 0.$$

The matrix $\mathbf{G}(\boldsymbol{\beta})$ is the nonrandom limit of $\mathbf{G}_N(\boldsymbol{\beta})$, that is,

$$\begin{aligned}\frac{\mathbf{G}_N(\boldsymbol{\beta})}{N} &\to \int_{\mathbb{R}^{p\times q}} \mathbf{Z}\mathbf{U}(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{U}'(\boldsymbol{\beta})\mathbf{Z}'\nu(d\mathbf{Z}) \\ &= \mathbf{G}(\boldsymbol{\beta})\end{aligned} \tag{20}$$

in probability as $N \to \infty$, where

$$\mathbf{U}(\boldsymbol{\beta}) = \left[\frac{\partial \mathbf{u}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right],$$

with $\boldsymbol{\eta} = \mathbf{Z}'\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ has generic elements

$$\sigma^{(ij)}(\boldsymbol{\beta}) = \begin{cases} -h_i(\mathbf{Z}'\boldsymbol{\beta})h_j(\mathbf{Z}'\boldsymbol{\beta}), & \text{if } i \neq j, \\ h_i(\mathbf{Z}'\boldsymbol{\beta})\big(1 - h_i(\mathbf{Z}'\boldsymbol{\beta})\big), & \text{if } i = j, \end{cases}$$

for $i, j = 1, \ldots, q$. Under some conditions, $\mathbf{G}(\boldsymbol{\beta})$ is a positive-definite matrix at the true parameter value and therefore its inverse exists.

As was noted in Fokianos and Kedem (1998), this approach is quite general and does not call for any Markov assumption. Previous related work on *conditional likelihood* estimation can be found in Fahrmeir and Kaufmann (1987) and Kaufmann (1987). The latter reference provides a rigorous treatment of consistency, asymptotic normality and efficiency of the maximum conditional likelihood estimator.

4.2 Testing Hypotheses

In applications, it is often necessary to test the general linear hypothesis

$$\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{against} \quad \mathrm{H}_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\beta}_0, \tag{21}$$

where $\mathbf{C}$ is an appropriate known matrix with full rank, say $r \leq p$. To this end, it is convenient to denote by $\tilde{\boldsymbol{\beta}}$ the restricted partial maximum likelihood estimator under the hypothesis (21). Then the most commonly used test statistics for testing (21) are:

- the partial likelihood ratio statistic

$$\lambda_N(\boldsymbol{\beta}) = -2\{l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\}; \tag{22}$$

- the Wald statistic

$$\begin{aligned} w_N(\boldsymbol{\beta}) = {} & \{\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\}' \\ & \cdot \{\mathbf{C}\mathbf{G}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}'\}^{-1}\{\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\}; \end{aligned} \tag{23}$$

- the partial score statistic

$$c_N(\boldsymbol{\beta}) = \mathbf{S}'_N(\tilde{\boldsymbol{\beta}})\mathbf{G}^{-1}(\tilde{\boldsymbol{\beta}})\mathbf{S}_N(\tilde{\boldsymbol{\beta}}). \tag{24}$$

The following theorem states the asymptotic distribution of these statistics (see also Fahrmeir, 1987).

THEOREM 4.2. *Under Assumption* A *of Fokianos and Kedem* (1998), *the test statistics* λ_N, w_N *and* c_N *are asymptotically equivalent*. *Furthermore*, *under the null hypothesis in* (21), *their asymptotic distribution is chi-square with* r *degrees of freedom*.

The behavior of all three test statistics under a sequence of alternatives is examined in Fahrmeir and Kaufmann (1987), where Theorem 4.2 is applied in testing the homogeneity and order of a Markov chain, structural change and independence of two parallel time series.

5. OTHER MODELS FOR CATEGORICAL TIME SERIES

There are several other models that have been proposed in the literature for the analysis of categorical time series. For instance, the so-called integer autoregressive models—introduced in a series of articles by McKenzie (1985, 1986, 1988), Al-Osh and Alzaid (1987) and Alzaid and Al-Osh (1990)—imitate the common AR structure (see Box, Jenkins and Reinsel, 1994) in the sense that the *thinning* operator is applied instead of common scalar multiplication.

The first attempt to introduce autoregressive and moving average models for discrete-valued time series data was made by Jacobs and Lewis (1978a, b). These authors consider the so-called discrete autoregressive moving average models—or DARMA models. In this section, we review the *mixture transition distribution* model which has been found useful in numerous applications.

5.1 The Mixture Transition Distribution Model

The *mixture transition distribution* (MTD) model was introduced in Raftery (1985a), extending previous work of Pegram (1980), as a parsimonious approach for the analysis of higher order Markov chains.

The mixture transition distribution model bypasses the problem of an exponentially increasing number of free parameters for a Markov chain by specifying the conditional probability of observing $X_t = i_0$ given the past as a linear combination of contributions from $X_{t-1}, \ldots, X_{t-p}$. More precisely, it is assumed that

$$\begin{aligned} &\mathrm{P}[X_t = i_0 | X_{t-1} = i_1, \ldots, X_{t-p} = i_p] \\ &\quad = \sum_{j=1}^{p} \lambda_j \mathrm{P}[X_t = i_0 | X_{t-j} = i_j] \\ &\quad = \sum_{j=1}^{p} \lambda_j q_{i_j i_0}, \end{aligned} \tag{25}$$

where $i_0, \ldots, i_p$ belong to $\{1, 2, \ldots, m\}$, $q_{i_j i_0}$ are elements of the $m \times m$ transition matrix $\mathbf{Q}$ and the vector of lag parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ satisfies

$$\sum_{j=1}^{p} \lambda_j = 1, \quad \lambda_j \geq 0,$$

so that the right-hand side of (25) is between 0 and 1. An alternative set of restrictions for $\boldsymbol{\lambda}$ is given by Raftery and Tavaré (1994).

Besides reducing considerably the number of parameters to $m(m-1) + (p-1)$, model (25) enjoys several properties. It can be shown that the limiting behavior of the MTD model is the same as that of the full parameterized higher order Markov chain (Raftery, 1985a; Adke and Deshmukh, 1988).

Various generalizations of the MTD model have been proposed. For example, Raftery (1985b) considers the multimatrix mixture transition distribution model called MTDg. The MTDg model uses a different transition matrix for each lag as follows:

$$\begin{aligned} &\mathrm{P}[X_t = i_0 | X_{t-1} = i_1, \ldots, X_{t-p} = i_p] \\ &\quad = \sum_{j=1}^{p} \lambda_j q_{i_j i_0}^{(j)}. \end{aligned} \tag{26}$$

Model (26) is less parsimonious than model (25) in the sense that it requires $m(m-1)+1$ additional parameters for each lag. However, it accommodates a dynamic relation between each lag and time period.

The work in Le, Martin and Raftery (1996) and, more recently, in Wong and Li (2000) extends definition (25) to arbitrary state space. The spatial MTD model is investigated by Raftery and Banfield (1991) and Berchtold (2001), and the double-chain Markov model is studied by Berchtold (1999).

Estimation of the parameters $\boldsymbol{\lambda}$ and q_{ij} of the mixture transition model (25) is accomplished by maximizing the log-likelihood (Raftery and Tavaré, 1994; Berchtold, 2001)

$$\sum_{i_0, \ldots, i_p = 1}^{m} n_{i_0, \ldots, i_p} \log\left(\sum_{j=1}^{p} \lambda_j q_{i_j i_0} \right)$$

subject to constraints for $\boldsymbol{\lambda}$. Here $n_{i_0, \ldots, i_p}$ counts the number of sequences $\{X_t = i_0, \ldots, X_{t-p} = i_p\}$. Alternative estimation methods include the minimum χ^2 estimation (Raftery and Tavaré, 1994) and the E-M algorithm (Le, Martin and Raftery, 1996). Software (MTD and GMTD) for fitting the mixture transition model as described above is available at http://lib.stat.cmu.edu/general. A thorough review of the mixture transition distribution model for higher order Markov chains and non-Gaussian time series can be found in Berchtold and Raftery (1999).

6. EXAMPLES

The regression methodology for categorical time series can be employed in diverse applications as this section illustrates in terms of DNA, soccer and sleep data. Moreover, models such as (4) and (8) offer great flexibility and accommodate dependence by inclusion of past values of the response and other covariates when available.

6.1 Explanatory Analysis of DNA Sequence Data

Regression models for categorical time series can be used in explanatory analysis of DNA sequence data as shown next by model fitting and testing, conditional on past response values.

A DNA sequence consists of four nucleotides differing only in the nitrogenous base, whose order determines the genetic information of each organism. The four nucleotides are given one-letter abbreviations as shorthand as follows:

- A is for adenine;
- G is for guanine;
- C is for cytosine;
- T is for thymine.

Adenine and guanine are purines—the larger of the two types of bases found in DNA—while cytosine and thymine are pyrimidines.

Thus, a strand of DNA can be represented as a sequence of letters from $\{A, C, G, T\}$ and can be viewed as a *nominal* categorical time series with the assignment $A = 1$, $C = 2$, $G = 3$ and $T = 4$. For more information, see Waterman (1995).

We present an explanatory analysis for DNA sequence data of the gene BNRF1 of the Epstein–Barr virus (see Shumway and Stoffer, 2000, Section 5.9) considering only the first 1000 observations—the whole data set is 3,954 long. The idea is to apply the multinomial logit model (4) by fitting a series of various-order models. For example, a first-order model is given by

$$\log\left(\frac{\pi_{ti}(\boldsymbol{\beta})}{\pi_{t4}(\boldsymbol{\beta})}\right) = \beta_{i0} + \beta_{i1}Y_{(t-1)1} + \beta_{i2}Y_{(t-1)2} + \beta_{i3}Y_{(t-1)3} \tag{27}$$

for $i = 1, 2, 3$ and is denoted by $1 + \mathbf{Y}_{t-1}$. A second-order model is labeled $1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2}$ and consists of (27) plus a linear combination in terms of $Y_{(t-2)1}$, $Y_{(t-2)2}$, $Y_{(t-2)3}$, and so on.

TABLE 1
Candidate models for the gene BNRF1 of the Epstein–Barr virus DNA sequence data

Model 1	$1 + \mathbf{Y}_{t-1}$
Model 2	$1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2}$
Model 3	$1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2} + \mathbf{Y}_{t-3}$
Model 4	$1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2} + \mathbf{Y}_{t-3} + \mathbf{Y}_{t-4}$

Table 1 lists the models applied to the DNA sequence data, and Table 2 reports the inferential results where the second column lists the number of estimated parameters, the third column reports the *deviance* of the model

$$D = -2\sum_{t=1}^{N}\sum_{j=1}^{m} Y_{tj}\log\pi_{tj}(\hat{\boldsymbol{\beta}}) \tag{28}$$

and the next two correspond to AIC and BIC criteria. The last two columns give the values of the likelihood ratio test statistic (22) together with its p-values for testing the order of the model. Thus, the value 13.48 is not significant, and therefore we might not include a lagged value of order 2 of the response. In other words, the hypothesis that $\mathbf{Y}_{t-2}$ should not enter the regression equation is accepted. Similarly, the p-value of 0.2121 casts a doubt on the inclusion of $\mathbf{Y}_{t-4}$, while a p-value of 0.0698 indicates that the inclusion of $\mathbf{Y}_{t-3}$ is reasonable. The last column is constructed by appealing to the chi-square distribution with 9 degrees of freedom. Throughout the analysis, we use $N = 996$ observations.

The results from Table 2 show that the AIC criterion is minimized for Model 1, while the BIC criterion is minimized for the independence model. However, the likelihood ratio test also indicates that an adequate model for the data at hand consists of an intercept, $\mathbf{Y}_{t-1}$ and $\mathbf{Y}_{t-3}$. For $1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-3}$, the number of estimated parameters is $p = 21$, $D = 2663.20$, BIC $=$ 2808.17, AIC $=$ 2705.20 and its p-value is 0.0968 when compared with Model 3.

TABLE 2
Comparison of different multinomial logit models for the gene BNRF1 of the Epstein–Barr virus DNA sequence data ($N = 996$)

Model	p	D	**AIC**	**BIC**	λ_N	p**-value**
Independence	3	2711.31	2717.31	2732.02		
Model 1	12	2677.75	2701.75	2760.60	33.56	0.0001
Model 2	21	2664.27	2706.27	2809.25	13.48	0.1420
Model 3	30	2648.41	2708.41	2855.52	15.86	0.0698
Model 4	39	2639.39	2714.39	2905.63	12.02	0.2121

TABLE 3
Estimated parameters β_{ij}, $i = 1, 2, 3$, $j = 0, \ldots, 6$, for model $1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-3}$ with their standard errors for the gene BNRF1 of the Epstein–Barr virus DNA sequence data

i	1	$\mathbf{Y}_{(t-1)1}$	$\mathbf{Y}_{(t-1)2}$	$\mathbf{Y}_{(t-1)3}$	$\mathbf{Y}_{(t-3)1}$	$\mathbf{Y}_{(t-3)2}$	$\mathbf{Y}_{(t-3)3}$
1	−0.908	0.541	0.665	1.071	0.534	0.167	0.787
	(0.110)	(0.171)	(0.080)	(0.059)	(0.160)	(0.084)	(0.056)
2	−0.438	0.423	0.288	0.904	0.558	0.486	0.784
	(0.100)	(0.150)	(0.074)	(0.055)	(0.150)	(0.076)	(0.053)
3	0.165	0.266	−0.412	0.584	0.262	0.320	0.422
	(0.097)	(0.141)	(0.075)	(0.054)	(0.146)	(0.074)	(0.053)

The estimated parameters for this model are given in Table 3 together with their standard errors in parentheses. The standard errors are computed after taking the square root along the diagonal of $\mathbf{G}_N^{-1}(\boldsymbol{\beta})$ which is approximated by $\mathbf{G}^{-1}(\boldsymbol{\beta})/N$; see (17) and (20), respectively. Figure 6 shows the sample autocorrelation plot of the squared Pearson residuals

$$\hat{r}_t = (\mathbf{Y}_t - \hat{\boldsymbol{\pi}}_t)' \hat{\boldsymbol{\Sigma}}_t^{-1} (\mathbf{Y}_t - \hat{\boldsymbol{\pi}}_t), \tag{29}$$

where $\hat{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t(\hat{\boldsymbol{\beta}})$, for $t = 1, \ldots, N$, suggesting that the fit is quite reasonable. Further evidence of this fact is manifested in Table 4, where the values of the power divergence statistic $\mathbf{I}_\lambda$ (see Read and Cressie, 1988; Fokianos, 2002) are tabulated for different λ. It is seen that all the $\mathbf{I}_\lambda$ values are less than 1.96 in absolute value, which confirms from another point of view that model $1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-3}$ is adequate.

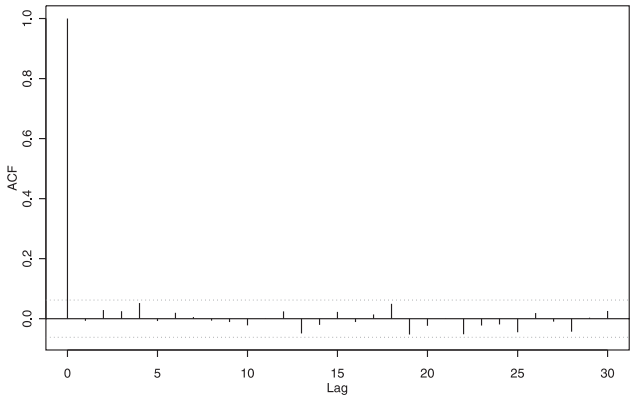

FIG. 6. *Sample autocorrelation function of the squared Pearson residuals corresponding to model $1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-3}$ for the gene BNRF1 of the Epstein–Barr virus DNA sequence data.*

The transition probabilities

$$\mathrm{P}(Y_t = i | Y_{t-1} = j, Y_{t-3} = l)$$

for $i, j, l = 1, 2, 3, 4$ are estimated by substitution of the maximum partial likelihood estimators into the regression equation of π_{ti} using (2). Table 5 reports the transition probabilities among the different states where, for example, if $Y_{t-3} = A$ and $Y_{t-1} = T$, then the transition probability to $Y_t = C$ is equal to 0.2592.

Table 6 reports the results of models (1), (25) and (26) and should be compared with Table 2. Notice that the number of reported parameters is different from $m^p(m-1)$ and $m(m-1) + (p-1)$ for the Markov chain and MTD models, respectively. The reason is that some transitions among the different states never occur.

The first line of Table 6 reports results under independence, the selected model under the BIC criterion. The next four rows show the analysis based on full Markov chain modeling, that is, model (1). We see that a Markov chain of order 4 leads to the smallest AIC with 321 parameters, while the BIC selects the first-order model with 12 parameters.

MTD fitting points to the first-order model. Indeed, an MTD model of order 1 is simply a Markov chain of order 1. It can be seen, though, that higher order MTD models do not affect the fit considerably since the

TABLE 4
Values of the power divergence statistic $\mathbf{I}_\lambda$ for model $1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-3}$ for the gene BNRF1 of the Epstein–Barr virus DNA sequence data

λ	−0.8	−0.4	−0.1	0.3	0.6	0.8	1	1.2
Value	−1.235	−1.056	−0.938	−0.774	−0.640	−0.556	−0.475	−0.396

TABLE 5
Estimated transition matrix from model $1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-3}$ *for the gene BNRF*1 *of the Epstein–Barr virus DNA sequence data*

		Y_{t-1}			
Y_{t-3}	Y_t	A	C	G	T
A	A	0.2004	0.2756	0.2352	0.1583
	C	0.2915	0.3097	0.3257	0.2592
	G	0.3389	0.2089	0.3219	0.3526
	T	0.1692	0.2058	0.1172	0.2299
C	A	0.1479	0.2107	0.1763	0.1149
	C	0.2889	0.3179	0.3279	0.2526
	G	0.3828	0.2443	0.3692	0.3916
	T	0.1804	0.2271	0.1266	0.2409
G	A	0.2167	0.2972	0.2511	0.1738
	C	0.3069	0.3251	0.3384	0.2770
	G	0.3342	0.2053	0.3135	0.3531
	T	0.1422	0.1724	0.0970	0.1961
T	A	0.1643	0.2289	0.2001	0.1249
	C	0.2335	0.2513	0.2705	0.1997
	G	0.3652	0.2279	0.3596	0.3656
	T	0.2370	0.2919	0.1688	0.3098

changes in deviance are rather small. Compared with the output of the multinomial logit model (4) for the DNA data, the MTD models reduce both the AIC and the BIC criteria. In addition, the number of parameters that need to be estimated is appreciably less than the number of parameters that need to be estimated for both the multinomial logit model and the full Markov chain. Note that, as in orders 2 and 3, the equality between the number of parameters for some models may not leave degrees of freedom for testing certain hypotheses.

The multilag MTDg model points to the first-order Markov chain. Notice again that, as with MTD of order 1, an MTDg model of order 1 is simply a Markov chain of order 1. Regarding the fitted MTDg models, here the number of parameters becomes large compared with those of the MTD model and this leads to an increase in both the AIC and the BIC values. Compared with the multinomial logit fit, the AIC and BIC values from the MTDg models are larger, except the order 1 model.

6.2 Soccer Forecasting

A popular weekly game in Greece is that of forecasting soccer game outcomes. Each week, a list of 13 soccer games is published by the Greek Organization of Forecasting Soccer Games in the form "Team A vs. Team B," where Team A plays at home. The 13 pairs vary every week. The published list usually consists of games played by the Greek First National League but occasionally some other games, either from the Greek Second National League or from a foreign league, enter the list. A potential bettor is challenged to forecast either 13, 12 or 11 correct outcomes by using the symbols "1" (Team A wins), "X" (a tie) or "2" (Team B wins).

The data consist of the true outcomes of the games for the first four positions of the list starting from 3/5/1995 and ending on 10/29/2000. This is a total of 289 sequential observations. Somewhat oddly, we record "X" as 3, and point out that there are some

TABLE 6
*Results from Markov chain and MTD models applied to gene BNRF*1 *of the Epstein–Barr virus DNA data* ($N = 996$)

Model	Number of Parameters	D	AIC	BIC
Independence	3	2711.31	2717.31	2732.02
Markov chain of order 1	12	2677.75	2701.75	2760.60
Markov chain of order 2	48	2627.68	2723.68	2959.06
Markov chain of order 3	179	2463.22	2821.22	3698.99
Markov chain of order 4	321	1808.33	2450.33	4024.44
MTD of order 1	12	2677.75	2701.75	2760.60
MTD of order 2	13	2677.75	2703.75	2767.51
MTD of order 3	13	2677.11	2703.11	2766.86
MTD of order 4	14	2676.18	2704.18	2772.83
MTDg of order 1	12	2677.75	2701.75	2760.60
MTDg of order 2	25	2664.27	2714.27	2836.87
MTDg of order 3	36	2647.27	2719.27	2895.80
MTDg of order 4	46	2631.12	2723.12	2948.70

TABLE 7
Frequencies for the soccer forecasting data

	"1"	"X"	"2"
Position 1	166	63	60
Position 2	150	71	68
Position 3	156	57	76
Position 4	155	63	71

weeks when the gambling game did not run on schedule. However, for data analysis purposes, we view these data as a regular *ordinal* time series with ordered categories "1," "X" and "2."

Table 7 reports the frequencies of the categories "1," "X" and "2." Thus, in the first position, among all the soccer games corresponding to the first four positions on the list, 166 games ended up a win for the home team, 63 games were a tie and 60 games were a loss. Figure 7 depicts time series plots of the first 150 outcomes for these data for the different positions. For each position, the weekly occurrences of the ordinal values "1," "X"(="3") and "2" define a categorical time series.

We investigate whether there is dependence among the games by analyzing these categorical time series. For each time series, we fit a proportional odds model (8) with lagged values of the response up to order 2 as covariates. The results—for $N = 287$—are summarized in Table 8. Let us consider the first-order model fitted for position 2. According to (8) and with suggestive notation,

$$\log\left[\frac{\mathrm{P}(Y_t \le \text{"1"}|\mathcal{F}_{t-1})}{\mathrm{P}(Y_t > \text{"1"}|\mathcal{F}_{t-1})}\right] = \theta_1 + \gamma_1 Y_{(t-1)1} + \gamma_2 Y_{(t-1)2},$$

$$\log\left[\frac{\mathrm{P}(Y_t \le \text{"X"}|\mathcal{F}_{t-1})}{\mathrm{P}(Y_t > \text{"X"}|\mathcal{F}_{t-1})}\right] = \theta_2 + \gamma_1 Y_{(t-1)1} + \gamma_2 Y_{(t-1)2}.$$

The corresponding estimators are $\hat{\theta}_1 = 0.251$, $\hat{\theta}_2 = 1.368$, $\hat{\gamma}_1 = -0.131$ and $\hat{\gamma}_2 = -0.415$ and their standard errors are 0.244, 0.257, 0.287 and 0.327, respectively.

From Table 8, the BIC criterion is minimized for the independence model for all positions, while the AIC criterion is minimized for the independence model with the single exception of position 2. Similarly, the deviance is not reduced significantly when entering the

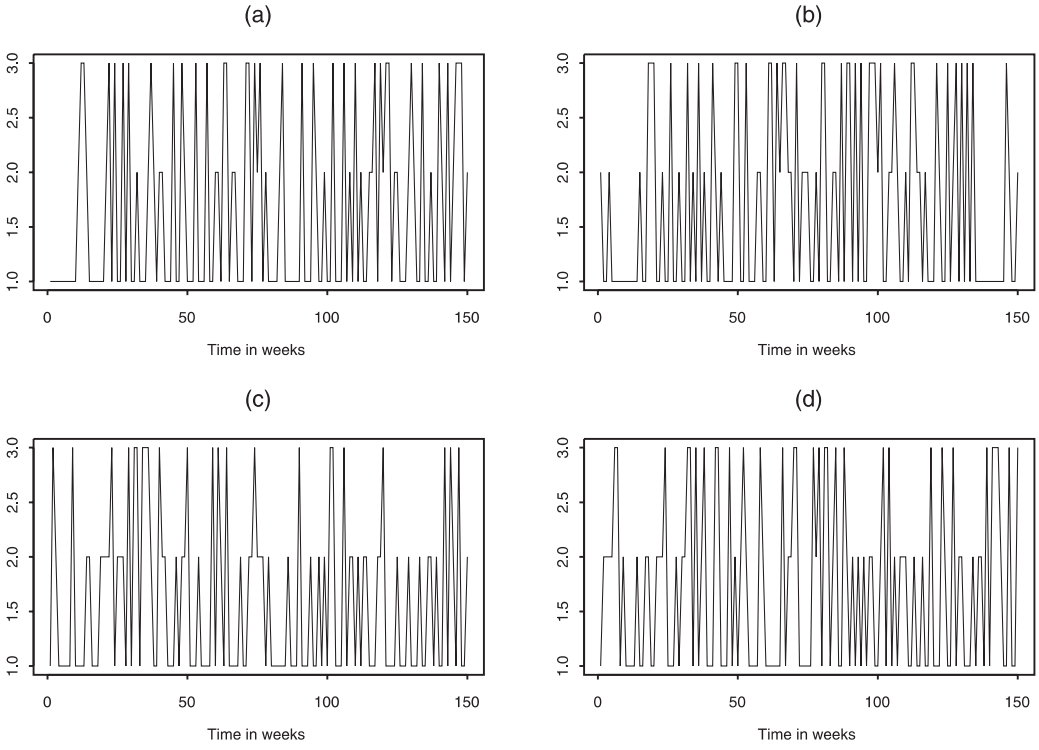

FIG. 7. *Time series plot of the first* 150 *observations for the soccer forecasting data*. (a) *Outcomes of first position*. (b) *Outcomes of second position*. (c) *Outcomes of third position*. (d) *Outcomes of fourth position*.

TABLE 8
Comparison of different proportional odds models for the soccer forecasting data

Time series	Model	p	D	AIC	BIC
Position 1	Independence	2	562.43	566.43	573.74
	$1 + \mathbf{Y}_{t-1}$	4	562.15	570.15	584.79
	$1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2}$	6	561.73	573.73	595.69
Position 2	Independence	2	588.63	592.63	599.95
	$1 + \mathbf{Y}_{t-1}$	4	586.84	594.84	609.47
	$1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2}$	6	578.09	590.09	612.05
Position 3	Independence	2	575.97	579.97	587.28
	$1 + \mathbf{Y}_{t-1}$	4	574.12	582.12	596.77
	$1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2}$	6	569.18	581.18	603.13
Position 4	Independence	2	580.33	584.33	591.65
	$1+\mathbf{Y}_{t-1}$	4	580.12	588.12	602.75
	$1+\mathbf{Y}_{t-1} + \mathbf{Y}_{t-2}$	6	579.93	591.93	613.88

lagged regressors into the model equation. It therefore seems reasonable to conclude that the independence model is quite adequate for the soccer forecasting data and that the betting game is fair. As expected from home games, "1" is more frequent than "X" and "2": "1" appears roughly 50% of the times, while the relative frequency of "X" and "2" is about 25% each.

We compare the obtained results with the fit from (1), (25) and (26) to the soccer forecasting data. Table 9 reports the results of this analysis only for the games played in the first position. We see that the proportional odds model performs better than all the alternatives considered in the table in the sense of minimizing both the AIC and the BIC. To explain this, notice the relatively small number of parameters required when fitting a proportional odds model. Furthermore, the results are consistent with the previous analysis. That is, the model of independence fits the soccer data quite well, leading once more to the conclusion that the soccer forecasting game is fair.

6.3 Sleep Data

The advantage of the regression models for categorical time series over the other models considered so far becomes more apparent when considering random time-dependent covariates.

The sleep data which have been discussed briefly—see Figure 1—consist of sleep state measurements of a newborn infant together with his heart rate (R_t) and temperature (T_t) sampled every 30 seconds. Recall that the sleep states are classified as:

(1) quiet sleep,
(2) indeterminate sleep,
(3) active sleep,
(4) awake.

The total number of observations is equal to 1024 and a plot of the data is displayed in Figure 8. The objective is to predict—or classify—the sleep state based on covariate information. In this respect, Figure 8 shows

TABLE 9
Results from Markov chain and MTD models applied to the soccer forecasting data for the first position ($N = 289$)

Model	Number of parameters	D	AIC	BIC
Independence	2	562.43	566.43	573.74
Markov chain of order 1	6	558.69	570.69	592.64
Markov chain of order 2	18	549.21	585.21	651.08
MTD of order 1	6	558.69	570.69	592.64
MTD of order 2	7	558.68	572.68	598.29
MTDg of order 1	6	558.69	570.69	592.64
MTDg of order 2	12	557.84	581.84	625.75

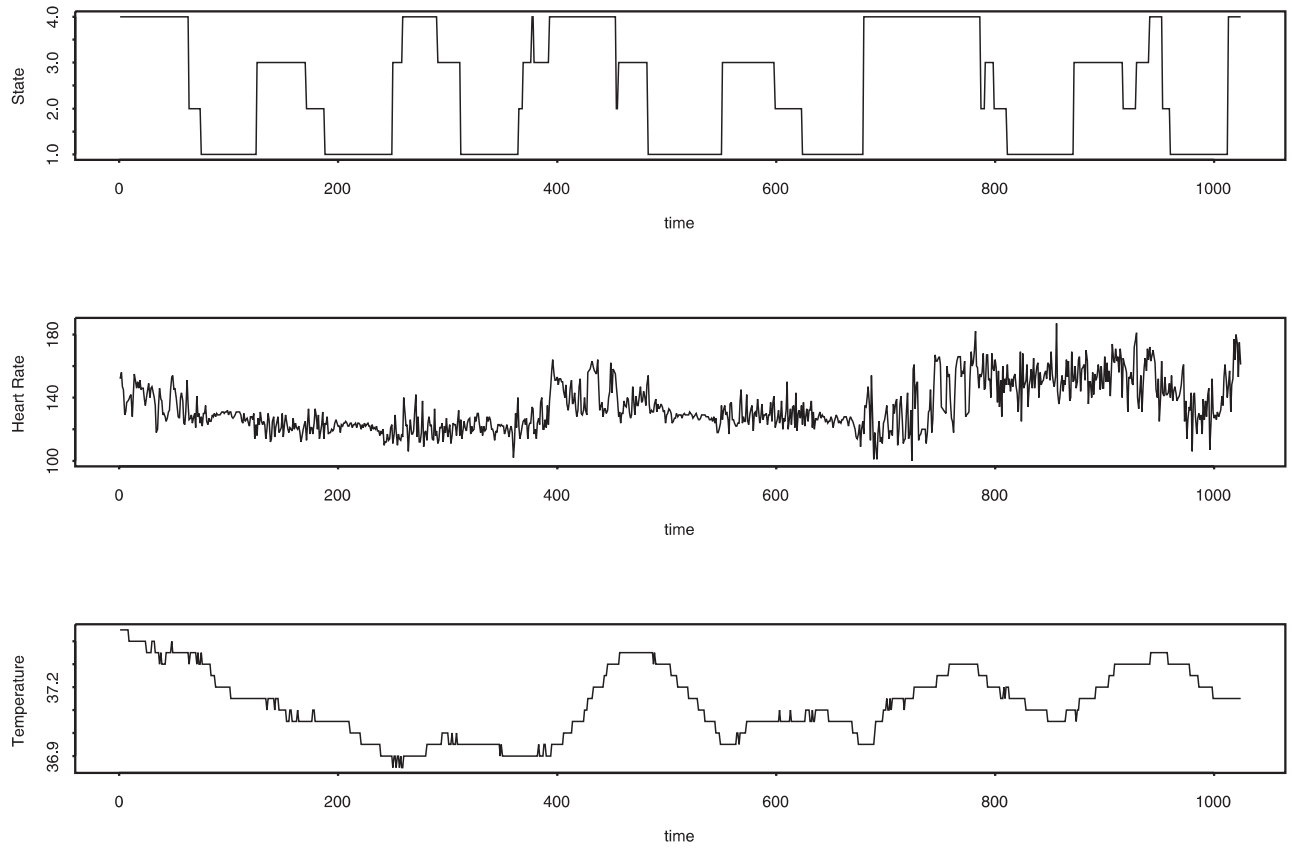

FIG. 8. *Time series plot for the sleep data* ($N = 1024$).

that sleep state depends on heart rate—higher values of heart rate tend to correspond to state (4).

To begin analyzing these data, notice that the response—the sleep state, say Y_t—is an ordered time series in the sense that "(4)" < "(1)" < "(2)" < "(3)"; that is, the response increases from awake to active sleep. By removing the first two observations, we fit several proportional odds models (8) to these data using only the first 700 observations. The remaining observations are used as a testing data set. The results of the analysis are summarized in Table 10.

Table 10 shows that a sensible model for the sleep data includes $\mathbf{Y}_{t-1}$ and the logarithm of heart rate ($\log R_t$). Comparing Model 1 with Model 7, we notice that $\mathbf{Y}_{t-1}$ is clearly a significant predictor. In addition, the deviance difference between Model 2 and Model 1 is 2.05 (p-value $= 0.1522$), suggesting that the logarithm of heart rate may be included in the model. Models 3, 4, 5 and 6 do not substantially enhance the fitted model, leading to the conclusion that temperature, and higher order lagged values of the response, are not significant predictors. These factors lead to the following

TABLE 10
Comparison of different proportional odds models for the sleep data ($N = 700$)

Model	Covariates	p	D	AIC	BIC
1	$1 + \mathbf{Y}_{t-1}$	6	389.56	401.56	428.86
2	$1 + \mathbf{Y}_{t-1} + \log R_t$	7	387.51	401.51	433.37
3	$1 + \mathbf{Y}_{t-1} + \log R_t + T_t$	8	387.32	403.32	439.73
4	$1 + \mathbf{Y}_{t-1} + T_t$	7	389.52	403.52	435.38
5	$1 + \mathbf{Y}_{t-1} + \mathbf{Y}_{t-2} + \log R_t$	10	387.28	407.28	452.79
6	$1 + \mathbf{Y}_{t-1} + \log R_{t-1}$	7	389.40	403.40	435.26
7	$1 + \log R_t$	4	1684.31	1692.31	1710.51

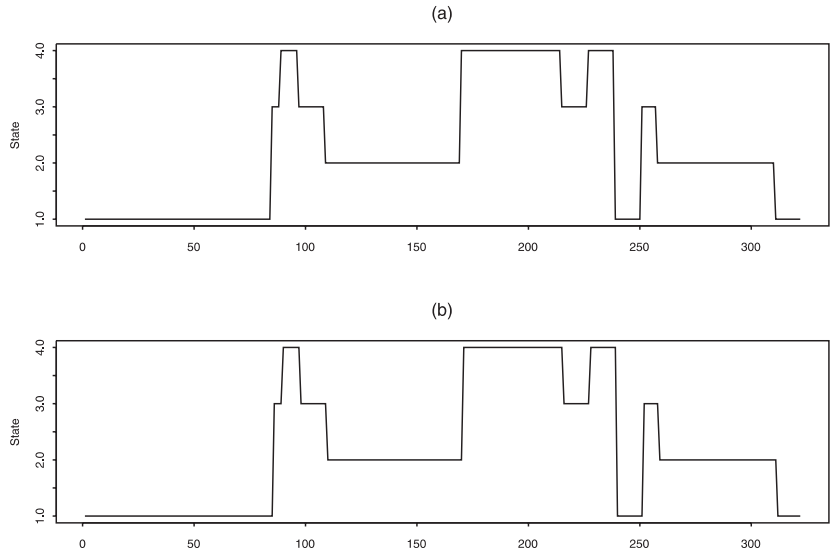

FIG. 9. (a) *Observed versus* (b) *predicted sleep states for model* $1 + \mathbf{Y}_{t-1} + \log R_t$ *applied to the testing data set* ($N = 322$).

model:

$$\log\left[\frac{\mathrm{P}(Y_t \le \text{“4”}|\mathcal{F}_{t-1})}{\mathrm{P}(Y_t > \text{“4”}|\mathcal{F}_{t-1})}\right] = \theta_1 + \gamma_1 Y_{(t-1)1} + \gamma_2 Y_{(t-1)2} + \gamma_3 Y_{(t-1)3} + \gamma_4 \log R_t,$$

$$\log\left[\frac{\mathrm{P}(Y_t \le \text{“1”}|\mathcal{F}_{t-1})}{\mathrm{P}(Y_t > \text{“1”}|\mathcal{F}_{t-1})}\right] = \theta_2 + \gamma_1 Y_{(t-1)1} + \gamma_2 Y_{(t-1)2} + \gamma_3 Y_{(t-1)3} + \gamma_4 \log R_t,$$

$$\log\left[\frac{\mathrm{P}(Y_t \le \text{“2”}|\mathcal{F}_{t-1})}{\mathrm{P}(Y_t > \text{“2”}|\mathcal{F}_{t-1})}\right] = \theta_3 + \gamma_1 Y_{(t-1)1} + \gamma_2 Y_{(t-1)2} + \gamma_3 Y_{(t-1)3} + \gamma_4 \log R_t,$$

with $\hat{\theta}_1 = -30.3529$, $\hat{\theta}_2 = -23.4931$, $\hat{\theta}_3 = -20.3495$, $\hat{\gamma}_1 = 16.7183$, $\hat{\gamma}_2 = 9.5338$, $\hat{\gamma}_3 = 4.7550$ and $\hat{\gamma}_4 = 3.5567$. The corresponding standard errors are 12.0517, 12.0128, 11.9858, 0.8726, 0.6306, 0.5018 and 2.4709.

Model 2 is applied to the testing data set which consists of 322 measurements. Figure 9 displays a time series plot of the observed versus predicted sleep states for the testing data set. The predicted responses are obtained by the following simple rule:

$$Y_t = j \quad \Longleftrightarrow \quad \max_k \hat{\pi}_{tk} = \hat{\pi}_{tj};$$

that is, category j is chosen if and only if its estimated transition probability is the maximum among the estimated transition probabilities. The misclassification rate of Model 2 is 0.034.

Some further diagnostics for Model 2 are given in Figure 10 and Table 11. Figure 10 displays cumulative periodogram plots (Priestley, 1981) for the Pearson

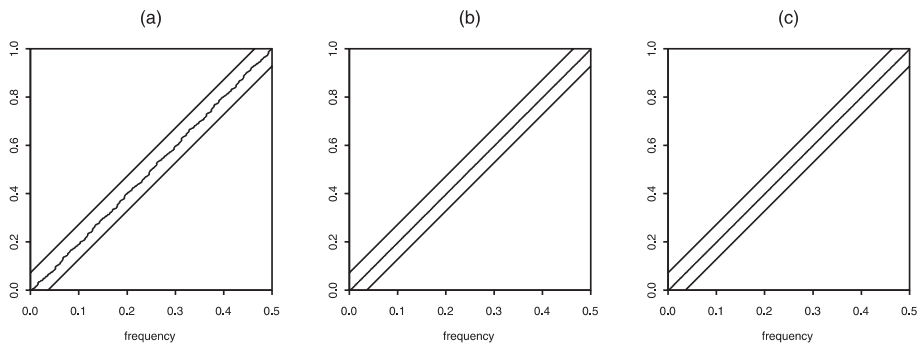

FIG. 10. *Cumulative periodogram plots for the Pearson residuals from model* $1 + \mathbf{Y}_{t-1} + \log R_t$ *applied to sleep data.*

TABLE 11
Values of the power divergence statistic $\mathbf{I}_\lambda$ *for model* $1 + \mathbf{Y}_{t-1} + \log R_t$ *applied to sleep data*

λ	-0.5	5	5.50	6	6.50	7
Value	8.813	4.581	2.767	1.667	1.002	0.601

residuals defined by

$$\hat{\mathbf{r}}_t = \hat{\mathbf{\Sigma}}_t^{-1/2}(\mathbf{Y} - \hat{\boldsymbol{\pi}}_t),$$

where $\hat{\mathbf{r}}_t$ is a q-dimensional vector, together with 95% confidence bands. Notice that for the sleep data $m = 4$, $q = 3$, so Figure 10(a) corresponds to the cumulative periodogram of the Pearson residuals for the first category and so on. In all these cases, we observe that the residual processes correspond to white noise. Table 11 lists the values of the power divergence statistic $\mathbf{I}_\lambda$ for different λ. We notice that for some λ the test is reassuring but this conclusion is not uniform. For an alternative approach to the problem of sleep state prediction using wavelet methods, see Nason, Sapatinas and Sawczenko (2001).

7. ADDITIONAL TOPICS

7.1 Alternative Modeling

Models for discrete-valued time series provide alternative approaches to categorical time series modeling. Important examples include higher order Markov chains (Azzalini, 1983; Raftery and Tavaré, 1994) and discrete autoregressive moving average (DARMA) models (Jacobs and Lewis, 1978a, b). Another useful class is that of variable-length Markov chains (VLMC) defined on a finite state space, where the Markov property is retained with variable order (Bühlmann and Wyner, 1999).

Another source of models are various transformations of an underlying process. Notable examples are categorical time series generated by "clipping" or "hard limiting" of a Gaussian process (Kedem, 1980, 1994). For an interesting extension of this to "discrete images" obtained by quantizing a Gaussian random field, see Kozintsev and Kedem (2000) and Kedem and Kozintsev (2000). Interestingly, under stationarity, parameters in the original series/field can be estimated quite effectively from the quantized data using very few (e.g., three) quantization levels. We mention Keenan (1982) as another example whereby a binary time series is generated according to an underlying strictly stationary but unobserved process. The connection between hidden Markov models and categorical time series has been explored in MacDonald and Zucchini (1997).

7.2 Spectral Analysis

Spectral analysis, a topic indigenous to time series, deserves serious consideration especially when the goal is to discover periodic components in the data (Priestley, 1981). Thus, we are led to consider the spectrum of a categorical time series. However, due to the qualitative nature of nominal data, the notion of spectrum is problematic. Recent work in this area by Stoffer, Tyler and McDougall (1993) and Stoffer, Tyler and Wendt (2000) attacks the problem by introducing the notions of scaling and assigning numerical values to the categories and that of the spectral envelope for selecting scales.

7.3 Longitudinal Data

Various authors have considered analysis of longitudinal categorical data. For a survey of results in this area, see the early article by Ashby et al. (1992) and the recent works by Pendergast et al. (1996), Agresti (1999) and Molenberghs and Lesaffre (1999). Inference for longitudinal multinomial data is mostly based on a generalized estimating equation approach (see Diggle, Liang and Zeger, 1994). Some key references include Stram, Wei and Ware (1988), where the authors suggest a method for comparing ordered categorical responses in two groups of subjects observed repeatedly allowing for time-dependent covariates and missing observations, and, more recently, Clayton (1992), Miller, Davis and Landis (1993), Williamson, Kim and Lipsitz (1995), Heagerty and Zeger (1996, 1998), Fahrmeir and Pritscher (1996) and Sutradhar and Kovacevic (2000). The work by Heagerty and Zeger (1998) proposes the *lorelogram* which can be used as a data analysis tool for exploring dependence in longitudinal categorical responses, while the contribution by Kosorok and Chao (1996) develops a Markov chain model for repeated ordinal data in continuous time.

ACKNOWLEDGMENTS

Thanks are due to Professor Casella, an Associate Editor and both reviewers whose constructive and helpful criticism improved the presentation. The soccer forecasting data were provided by the Greek Organization for Forecasting Soccer Games, while the sleep data were provided by F. Sapatinas.

REFERENCES

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

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York.

AGRESTI, A. (1999). Modeling ordered categorical data: Recent advances and future challenges. *Statistics in Medicine* **18** 2191–2207.

AL-OSH, M. A. and ALZAID, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *J. Time Ser. Anal.* **8** 261–275.

ALZAID, A. A. and AL-OSH, M. A. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *J. Appl. Probab.* **27** 314–324.

ASHBY, M., NEUHAUS, J., HAUCK, W., BACCHETTI, P., HEILBRON, D., JEWELL, N., SEGAL, M. and FUSARO, R. (1992). An annotated bibliography of methods for analyzing correlated categorical data. *Statistics in Medicine* **11** 67–99.

AZZALINI, A. (1983). Maximum likelihood estimation of order m for stationary stochastic process. *Biometrika* **70** 381–387.

BASAWA, I. V. and PRAKASA RAO, B. L. S. (1980). *Statistical Inference for Stochastic Processes*. Academic Press, London.

BERCHTOLD, A. (1999). The double chain Markov model. *Comm. 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. 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. Statist., Univ. Washington, Seattle.

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

BOX, G. P., JENKINS, G. M. and REINSEL, G. C. (1994). *Time Series Analysis, Forecasting and Control,* 3rd. ed. Prentice-Hall, Englewood Cliffs, NJ.

BRILLINGER, D. R. (1996). An analysis of an ordinal-valued time series. In *Athens Conference on Applied Probability and Time Series, II: Time Series Analysis. Lecture Notes in Statist.* **115** 73–87. Springer, New York.

BRILLINGER, D. R., MORETTIN, P. A., IRIZARRY, R. A. and CHIANN, C. (2000). Some wavelet-based analyses of Markov chain data. *Signal Processing* **80** 1607–1627.

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

CLAYTON, D. G. (1992). Repeated ordinal measurements: A generalized estimating equation approach. Technical report, Medical Research Council Biostatistics Unit, Cambridge, UK.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

DIGGLE, P. J., LIANG, K.-Y. and ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford Univ. Press, New York.

FAHRMEIR, L. (1987). Asymptotic testing theory for generalized linear models. *Statistics* **18** 65–76.

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

FAHRMEIR, L. and PRITSCHER, L. (1996). Regression analysis of forest damage by marginal models for correlated ordinal responses. *Environ. Ecol. Stat.* **3** 257–268.

FAHRMEIR, L. and TUTZ, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed. Springer, New York.

FOKIANOS, K. (2002). Power divergence family of tests for categorical time series models. *Ann. Inst. Statist. Math.* **54** 543–564.

FOKIANOS, K. and KEDEM, B. (1998). Prediction and classification of non-stationary categorical time series. *J. Multivariate Anal.* **67** 277–296.

FOKIANOS, K., KEDEM, B. and SHORT, D. (1996). Predicting precipitation level. *J. Geophys. Res. D*: *Atmospheres* **101** 26,473–26,477.

GUTTORP, P. (1995). *Stochastic Modelling of Scientific Data*. Chapman and Hall, London.

HEAGERTY, P. J. and ZEGER, S. L. (1996). Marginal regression models for clustered ordinal measurements. *J. Amer. Statist. Assoc.* **91** 1024–1036.

HEAGERTY, P. J. and ZEGER, S. L. (1998). Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *J. Amer. Statist. Assoc.* **93** 150–162.

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.

JOHNSON, V. E. and ALBERT, J. H. (1999). *Ordinal Data Modeling*. Springer, New York.

KARLIN, S. and TAYLOR, H. M. (1975). *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York.

KAUFMANN, H. (1987). Regression models for nonstationary categorical time series: Asymptotic estimation theory. *Ann. Statist.* **15** 79–98.

KEDEM, B. (1980). *Binary Time Series*. Dekker, New York.

KEDEM, B. (1994). *Time Series Analysis by Higher Order Crossings*. IEEE Press, New York.

KEDEM, B. and KOZINTSEV, B. (2000). Graphical bootstrap. In *Proc. Section on Statistics and the Environment* 30–32. Amer. Statist. Assoc., Alexandria, VA.

KEENAN, D. M. (1982). A time series analysis of binary data. *J. Amer. Statist. Assoc.* **77** 816–821.

KOSOROK, M. R. and CHAO, W.-H. (1996). The analysis of longitudinal ordinal response data in continuous time. *J. Amer. Statist. Assoc.* **91** 807–817.

KOZINTSEV, B. and KEDEM, B. (2000). Generation of "similar" images from a given discrete image. *J. Comput. Graph. Statist.* **9** 286–302.

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

LUCE, R. D. (1959). *Individual Choice Behavior*. Wiley, New York.

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

MCCULLAGH, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 109–142.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

McFADDEN, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (P. Zarembka, ed.) 105–142. Academic Press, New York.

McKENZIE, E. (1985). Some simple models for discrete variate time series. *Water Res. Bull.* **21** 645–650.

McKENZIE, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Adv. in Appl. Probab.* **18** 679–705.

McKENZIE, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Adv. in Appl. Probab.* **20** 822–835.

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.

MILLER, M. E., DAVIS, C. S. and LANDIS, J. R. (1993). The analysis of longitudinal polytomous data: Generalized estimated equations and connections with weighted least squares. *Biometrics* **49** 1033–1044.

MOLENBERGHS, G. and LESAFFRE, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine* **18** 2237–2255.

NASON, G. P., SAPATINAS, T. and SAWCZENKO, A. (2001). Wavelet packet modeling of infant sleep state using heart rate data. *Sankhyā Ser. B* **63** 199–217.

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

PENDERGAST, J. F., GANGE, S. J., LINDSTROM, M. J., NEWTON, M. A., PALTA, M. and FISHER, M. R. (1996). A survey of methods for analyzing clustered binary response data. *Internat. Statist. Rev.* **64** 89–118.

PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

PRUSCHA, H. (1993). Categorical time series with a recursive scheme and with covariates. *Statistics* **24** 43–57.

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. 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.

READ, T. R. C. and CRESSIE, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York.

SHUMWAY, R. H. and STOFFER, D. S. (2000). *Time Series Analysis and Its Applications*. Springer, New York.

SLUD, E. V. and KEDEM, B. (1994). Partial likelihood analysis of logistic regression and autoregression. *Statist. Sinica* **4** 89–106.

SNELL, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics* **20** 592–607.

STOFFER, D. S., TYLER, D. E. and McDOUGALL, A. J. (1993). Spectral analysis of categorical time series: Scaling and the spectral envelope. *Biometrika* **80** 611–622.

STOFFER, D. S., TYLER, D. E. and WENDT, D. A. (2000). The spectral envelope and its applications. *Statist. Sci.* **15** 224–253.

STRAM, D. O., WEI, L. J. and WARE, J. H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *J. Amer. Statist. Assoc.* **83** 631–637.

SUTRADHAR, B. C. and KOVACEVIC, M. (2000). Analysing ordinal longitudinal survey data: Generalised estimating equations approach. *Biometrika* **87** 837–848.

WATERMAN, M. S. (1995). *Introduction to Computational Biology*: *Maps*, *Sequences and Genomes*. Chapman and Hall, New York.

WILLIAMSON, J. M., KIM, K. M. and LIPSITZ, S. R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *J. Amer. Statist. Assoc.* **90** 1432–1437.

WONG, C. S. and LI, W. K. (2000). On a mixture autoregressive model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 95–115.

WONG, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14** 88–123.