AStA Adv Stat Anal (2015) 99:337–365
DOI 10.1007/s10182-014-0243-3

ORIGINAL PAPER

Coherent forecasting for stationary time series of discrete data

Raju Maiti · Atanu Biswas

Received: 15 August 2013 / Accepted: 24 December 2014 / Published online: 28 January 2015

Abstract Coherent forecasting for discrete-valued stationary time series is considered in this article. In the context of count time series, different methods of coherent forecasting such as median forecasting and mode forecasting are used to obtain h-step ahead coherent forecasting. However, there are not many existing works in the context of categorical time series. Here, we consider the case of a finite number of categories with different possible models, such as the Pegram's operator-based ARMA(p,q) model, the mixture transition distribution model and the logistic regression model, and study their h-step ahead coherent forecasting. Some theoretical results are derived along with some numerical examples. To facilitate comparison among the three models, we use some forecasting measures. The procedure is illustrated using one real-life categorical data, namely the infant sleep status data.

Keywords Pegram's model · Markov model · MTD model · Logistic regression model · Coherent forecasting

1 Introduction

Discrete-valued time series can broadly be classified into two categories, namely the count time series and the categorical time series. Categorical time series can again be of ordinal or nominal type. Some examples of count time series are the annual counts of hurricanes, the number of patients treated each day in an emergency department or the daily counts of swine flu cases in Mexico. Sleep status in successive minutes

R. Maiti · A. Biswas (✉)
Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, India
e-mail: atanu@isical.ac.in

R. Maiti
e-mail: rajumaiti@gmail.com

is one example of ordinal categorical time series. On the other hand, a sequence of rainfall data in which successive days are recorded as "wet" or "dry" is one example of nominal categorical time series.

This paper is concerned about the coherent forecasting of discrete-valued time series, i.e., for data which are discrete in nature. By coherent forecasting, we mean that the forecasting values are either integer or categorical. In the count time series context, very few works are available in modeling as well as for coherent forecasting. Freeland and McCabe (2004) discussed some methods of coherent forecasting for thinning operator-based Poisson integer-valued autoregressive model of order 1 [denoted by PINAR(1)], which was introduced in McKenzie (1985) and Al-Osh and Alzaid (1987). Later, this thinning-based INAR(1) model was extended to INAR(p), INMA(q) and INARMA(p, q) models by McKenzie (1988) and Alzaid and Al-Osh (1990). Although the h-step ahead conditional mean to make h-step ahead forecasting can be derived without knowing the exact h-step ahead forecasting distribution, in general this conditional mean may not be an integer and hence it is not coherent. Also for nominal categorical time series, where one cannot assign numerical values to the categories, conditional mean does not make any sense and hence cannot be used for forecasting purpose. However, many authors obtained the exact expression for h-step ahead forecasting distribution and used its median and mode, which are coherent by its nature, to study the h-step ahead coherent forecasting. Later Jung and Tremayne (2006), Bu and McCabe (2008) and Silva et al. (2009) also used the same methods to study the coherent forecasting in more general setup. But, in general, these models are not applicable in modeling categorical time series with finite number of categories.

Jacobs and Lewis (1978a, b, c), in a series of papers introduced a simple method for obtaining a stationary sequence of dependent random variables with a specified marginal distribution and correlation structure chosen independently. It was perhaps the first attempt to obtain a general class of simple models for discrete variate time series including categorical processes. These models are structurally based on the well-known autoregressive-moving-average processes and are referred to as DARMA models. However, the most well-known approach towards fitting categorical time series data is perhaps the mixture transition distribution (MTD) model, a class of models based on time homogeneous higher order Markov chain, proposed by Raftery (1985). Later it had been modified and generalized by Berchtold and Raftery (2002) and references therein. In contrast, Pegram (1980) used a very special kind of Markovian model towards fitting discrete-valued time series, especially for categorical time series. It is important to note that the model proposed by Pegram (1980) is equivalent to the DAR(p) model considered by Jacobs and Lewis (1978c, 1983). In particular, the DAR(1) process in Jacobs and Lewis (1978c) is exactly same as that of the Pegram's AR(1) process. In the recent past, Biswas and Song (2009) had extended the Pegram's autoregressive model of order p, denoted by PAR(p), to more general setup—Pegram's autoregressive and moving-average model [denoted by PARMA(p,q)] which is equivalent to the NDARMA(p, q) model of Jacobs and Lewis (1983). Also see the alternative representation of the model in Weiß and Göb (2008). Regression model for categorical time series was also developed and applied in sleep status data by Fokianos and Kedem (2003).

In this article, we derive the exact h-step ahead coherent forecasting distributions of three discrete time series models, namely PARMA(p, q), MTD model of order p or MTD(p) and logistic regression model of order p or Logistic(p). It is important to note that, if a categorical time series has $k + 1$ categories, then the number of parameters to be estimated in the PARMA(p, q) model is only $(k + p + q)$, whereas it is $(k(k + 1) + p - 1)$ for the MTD(p) model and pk^2 for the Logistic(p) model. In other words, the PARMA models involve much less number of parameters compared to the other two models for sufficiently large values of k and p. In addition, the PARMA models exhibit the classical Yule–Walker serial dependence structure and it carries simple stochastic properties such as stationarity, ergodicity and so on. However, the model has one big disadvantage that it can only be used for time series exhibiting long runs of a certain value. In spite of the limitation, it is evident that the PARMA models are more flexible in terms of the range of correlation and the ease of interpretation. Therefore, in this article, forecasting study for the PARMA(p,q) model is carried out in detail with MTD and logistic models. Different methods of coherent forecasting for ordinal and nominal categorical time series, e.g., median and mode predictors are discussed. To study the forecasting performance, different measures of forecasting accuracy are studied. The list includes percentage of true prediction, Kolmogorov–Smirnov distance, Euclidean distance, maximum absolute distance between true and predicted distributions. In addition, we introduce a different notion of interval forecasting based on highest predicted probability (HPP), namely $100(1 - \alpha)$ % HPP set, and study its performance using some simulation studies. All these methods are illustrated using one real dataset of ordinal categorical time series, namely infant sleep status data.

The rest of the article is organized as follows. In Sect. 2, different methods of coherent forecasting with some measures of forecasting accuracy are discussed to study the forecasting performance. Coherent forecasting for PAR(p), PMA(q) and PARMA(p, q) models is presented in Sect. 3. Coherent forecasting for MTD(p) and Logistic(p) models is discussed in Sects. 4 and 5, respectively. Some extensive simulation results are presented in Sect. 6. In Sect. 7, a practical categorical data, namely infant sleep status data, are analyzed to illustrate the proposed methods. Section 8 concludes. All technical proofs are relegated to the Appendix.

2 Coherent forecasting

It is important to note that, forecasting which is an integral part of time series analysis, has received very little attention in the discrete-valued time series literature, especially in categorical time series analysis. In the context of count time series, Freeland and McCabe (2004) have introduced some coherent methods of h-step ahead forecasting. The list includes nearest integer of mean predictor, median predictor and mode predictor. If the time series data are categorical, then the nearest integer of mean predictor cannot be used since moments are not defined there. To use median predictor for categorical time series, the order of the categories is mandatory and hence median predictor can only be used for ordinal/ordered categorical time series. However, mode predictor does not depend on the order of the categories, and hence can always be used to obtain the h-step ahead coherent forecasting.

On the other hand, to examine the forecasting accuracy for time series of real-valued data, one can always use the popular measures like predicted root mean squared error (PRMSE) or predicted mean absolute error (PMAE) which can be defined as follows. Let $\{Y_t\},\ t = 1, 2, \ldots, N$ be a time series and let us denote $\mathcal{Y}_n = \{Y_n, Y_{n-1}, \ldots, Y_1\}$, then

$$\begin{aligned} \text{PRMSE}(h) &= \sqrt{E\left(\left(Y_{n+h} - \widehat{Y}_{n+h}\right)^2 |\mathcal{Y}_n\right)}; \qquad h = 1, 2, \ldots \\ &\hat{=} \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\widehat{y}_{(n+h)i} - y_{(n+h)i}\right)^2}, \end{aligned}$$

$$\begin{aligned} \text{PMAE}(h) &= E\left(\left|Y_{n+h} - \widehat{Y}_{n+h}\right| |\mathcal{Y}_n\right); \qquad h = 1, 2, \ldots \\ &\hat{=} \frac{1}{M}\sum_{i=1}^{M}\left|\widehat{y}_{(n+h)i} - y_{(n+h)i}\right|. \end{aligned}$$

where $y_{(n+h)i}$ be the true ith observation at time point $(n + h)$ and $\widehat{y}_{(n+h)i}$ be the predicted observation at the same time point observed by some forecasting methods and M is the number of iterations.

Unlike for time series of real-valued data, the PRMSE and PMAE cannot be observed, particularly for nominal categorical time series. For ordinal categorical process, although these measures can be observed after assigning some numbers to the categories, but these may lead to some wrong conclusions since there is a no unique way to assign numbers to the ordinal categories (discussed earlier). However, to examine the forecasting accuracy for count and categorical data, we can always use measure like percentage of true prediction (PTP) which is defined as

$$\begin{aligned} \text{PTP}(h) &= E\left(I\left(Y_{n+h} = \widehat{Y}_{n+h}\right)|\mathcal{Y}_n\right) \times 100; \qquad h = 1, 2, \ldots \\ &\hat{=} \frac{1}{M}\sum_{i=1}^{M} I\left(y_{(n+h)i} = \widehat{y}_{(n+h)i}\right) \times 100. \end{aligned}$$

In addition, we intend to propose some popular distance functions between true and predicted distributions as the measures of forecasting accuracy to study the forecasting accuracy for categorical time series analysis. The list includes (discrete) Kolmogorov–Smirnov distance (KSD), Euclidean distance (ED) (see, e.g., Carruth et al. 2012), and maximum absolute difference (MAD) which are defined as follows.

Let $\{Y_t\},\ t = 1, 2, \ldots, N$ be a time series of categorical data with $(k + 1)$ many categories $\{C_0, C_1, \ldots, C_k\}$, and let us assume that $\mathbf{p}_h = (p_h(0), p_h(1), \ldots, p_h(k))$ denotes the h-step ahead true distribution of Y_{n+h} given $\mathcal{Y}_n$ with $\sum_{i=0}^{k} p_h(i) = 1$, where $p_h(i)$ denotes the probability mass function of Y_{n+h} at C_i given $\mathcal{Y}_n$. Let $\widehat{\mathbf{p}}_h$ denote the h-step ahead forecasting distribution, then KSD, ED and MAD functions can be defined as

$$\mathrm{KSD}(\mathbf{p}_h, \widehat{\mathbf{p}}_h) = \max_{0 \le i \le k} \left| \sum_{j=0}^{i} p_h(j) - \sum_{j=0}^{i} \widehat{p}_h(j) \right|,$$

$$\mathrm{ED}(\mathbf{p}_h, \widehat{\mathbf{p}}_h) = \sqrt{\sum_{j=0}^{k} (p_h(j) - \widehat{p}_h(j))^2},$$

and

$$\mathrm{MAD}(\mathbf{p}_h, \widehat{\mathbf{p}}_h) = \max_{0 \le j \le k} |p_h(j) - \widehat{p}_h(j)|.$$

It is important to mention that unlike KSD, the other two measures can be applied to any type of categorical time series—nominal or ordinal. However, KSD which is the maximum absolute difference between cumulative distribution functions depends on the ordering of the categories. Thus, when there is a natural ordering of the data, KSD is recommended, while the ED and MAD are more reliable and more easily understood than the KSD when there is no natural ordering (or partial order). In the context of goodness of fit of categorical data analysis, a comparison study between ED and KSD is also available in Carruth et al. (2012).

As far as the interval forecasting for categorical time series process is concerned, especially for nominal time series, it is not feasible to obtain the usual prediction interval of Y_{n+h} given $\mathcal{Y}_n$. However, we can use some notion of prediction set in place of prediction interval, e.g., highest predicted probability (HPP) set which is defined as follows:

Definition A $100(1-\alpha)\,\%$ HPP set of Y_{n+h} given $\mathcal{Y}_n$, denoted by $\mathcal{S}_h$ and is defined as

$$\mathcal{S}_h = \{C_j,\ j \in J :\ p_h(j) \ge k_\alpha\}$$

where $J = \{0, 1, \ldots, k\}$ and k_α is the largest number such that

$$P(Y_{n+h} \in \mathcal{S}_h | \mathcal{Y}_n) = \sum_{\{j : C_j \in \mathcal{S}_h\}} p_h(j) \ \ge (1-\alpha).$$

Based on the above definition, we can obtain the $100(1-\alpha)\,\%$ HPP set, $\mathcal{S}_h$, of Y_{n+h} given $\mathcal{Y}_n$. It is important to notice that $\mathcal{S}_h$ does not depend on the nature of the categories, and the usual length of $\mathcal{S}_h$ (like the length of the prediction interval) does not make sense here. Therefore, we introduce a notion of length of $\mathcal{S}_h$, namely the cardinality of $\mathcal{S}_h$, denoted by $n(\mathcal{S}_h)$ which gives the number of elements in the set, and study its behavior using some simulation studies to obtain the interval forecasting accuracy against h in the later sections.

3 Coherent forecasting for Pegram's operator-based ARMA(p, q) models

3.1 Pegram's operator

Pegram's operator $*$, when operated on U and V, say, defines a new random variable Z as a mixture of U and V with mixing coefficients ϕ and $1-\phi$. This is defined as

$$Z = (U, \phi) * (V, 1-\phi), \tag{3.1}$$

where the marginal probability function of Z is given by

$$P(Z=j) = \phi P(U=j) + (1-\phi)P(V=j), \quad j = 0, 1, \ldots$$

The mixing operator $*$ can be easily extended to handle more than two discrete variables. Pegram's (1980) construction has been extended to ARMA(p,q) model by Biswas and Song (2009) and Biswas and Guha (2009). The extension is equivalent to the NDARMA model by Jacobs and Lewis (1983). Also an alternative representation of the NDARMA model is available in Weiß and Göb (2008). The key advantage of Pegram's operator is that it provides a flexible mixing operation that enables us to define the mixture among a finite number of probability distributions of categorical random variables. It may be noted here that in this model the value of the variable of interest at time t depends on its value at time $(t-1)$ only through the probability of being equal to it and so on, as pointed out by Raftery (1985), who argued that the dependence patterns for such models are restricted.

3.2 Pegram's operator-based AR(p) model

Based on the above mixing operator $*$, Pegram (1980) constructed a stationary AR(p) process. Let $\{Y_t\}$ denote the response series with $(k+1)$ categories $\{C_0, C_1, \ldots, C_k, \}$. Then the process $\{Y_t\}$ is defined as

$$Y_t = (I(Y_{t-1}), \phi_1) * (I(Y_{t-2}), \phi_2) * \cdots * (I(Y_{t-p}), \phi_p) * (\epsilon_t, 1-\phi_1-\phi_2-\cdots-\phi_p), \tag{3.2}$$

which is a mixture of $(p+1)$ discrete distributions, where $P(\epsilon_t = C_i) = p_i$, $i = 0, 1, \ldots, k$, and it is denoted by $\epsilon_t \sim D((C_i, p_i), i = 0, 1, \ldots, k)$, with respective mixing weights being $\phi_1, \ldots, \phi_p$ with $\phi_i \in (0,1)$, $i = 1, \ldots, p$, and $\sum_{i=1}^{p} \phi_i \in (0,1)$. For every $t = 0, \pm 1, \pm 2, \ldots$ the conditional probability function takes the form

$$\begin{aligned} P(Y_t = C_i | Y_{t-1} = C_{i_1}, \ldots, Y_{t-p} = C_{i_p}) &= \phi_1 I(i_1 = i) + \cdots + \phi_p I(i_p = i) \\ &\quad + (1-\phi_1-\phi_2-\cdots-\phi_p) p_i, \end{aligned} \tag{3.3}$$

where ϕ_j, $j = 1, \ldots, p$, are chosen such that the polynomial equation $1-\phi_1 z - \cdots - \phi_p z^p = 0$ has roots lying outside of the unit disc. Here $I(\cdot)$ is the indicator function such that $I(A) = 1$ or 0 whether A occurs or not.

Taking expectation in both sides of (3.3), we observe that $P(Y_{t-h} = C_i) = p_i$ for $h = 1, \ldots, p$, resulting in $P(Y_t = C_i) = p_i$, which implies the marginal stationarity, i.e., marginally $Y_t \sim D((C_i, p_i), i = 0, 1, \ldots, k)$ for all t.

For a stationary PAR(1) model the following simple Theorem is proved in Biswas and Song (2009).

Theorem 1 *For $h \geq 1$, we have*

$$P(Y_{t+h} = C_i | Y_t = C_j) = \phi^h I(j = i) + (1 - \phi^h) p_i. \tag{3.4}$$

A more general result for the NDARMA(p, q) model, which is equivalent to the PARMA(p, q) model, was derived by (Weiß and Göb, 2008, Section 5), although the transition probability distribution for $h > 1$ was not derived there.

It is important to mention that, if the time series is categorical, especially nominal categorical, where one cannot assign numerical values to the categories, the moments, autocorrelation function cannot be defined. Although the autocorrelation function is not defined, some measures of serial association can always be defined for such processes. In the recent past, Weiß and Göb (2008) proposed several measures of association in the context of modeling categorical time series. The list includes popular measures like Goodman and Kruskal's τ, Goodman and Kruskal's λ, Cramer's ν, Cohen's κ and many others (see Weiß and Göb 2008 for details). These measures can also be used to select the order of the models. Even if the categories are ordinal type where one can assign some ordered numerical scalings, the above measures can also be used as alternatives to the autocorrelation. This is because different people using their own numerical scalings will get different values of moments and autocorrelation for the same categorical time series. Based on these measures, a detailed numerical study is carried out in latter sections.

Now to study the different notions of h-step ahead coherent forecasting and different measures of forecasting accuracy discussed in earlier section, we derive the following results.

Theorem 2 *For a stationary PAR(p) model, the h-step ahead forecasting distribution of Y_{n+h} given $\mathcal{Y}_n$ is given by*

$$\begin{aligned} p_h(i; \boldsymbol{\phi}) &= P(Y_{n+h} = C_i | \mathcal{Y}_n) \\ &= \eta_{h1} I(Y_n = C_i) + \cdots + \eta_{hp} I(Y_{n-p+1} = C_i) + (1 - \eta_{h1} - \cdots - \eta_{hp}) p_i \\ &= \boldsymbol{\eta}_h^T \boldsymbol{e} + \left(1 - \boldsymbol{\eta}_h^T \mathbf{1}\right) p_i, \end{aligned} \tag{3.5}$$

where the vector of h-step ahead parameters $\boldsymbol{\eta}_h = \left(\eta_{h1}, \eta_{h2}, \ldots, \eta_{hp}\right)^T$ is given by

$$\boldsymbol{\eta}_h = \boldsymbol{\Phi}^{h-1} \boldsymbol{\phi}, \tag{3.6}$$

with

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & \phi_2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & \cdots & \phi_{p-1} & 1 \\ 0 & 0 & 0 & \cdots & 0 & \phi_p \end{pmatrix}, \boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} I(Y_n = C_i) \\ I(Y_{n-1} = C_i) \\ \vdots \\ I(Y_{n-p+1} = C_i) \end{pmatrix},$$

and $\boldsymbol{\Phi}^{h-1} = \underbrace{\boldsymbol{\Phi} \times \boldsymbol{\Phi} \times \cdots \times \boldsymbol{\Phi}}_{h-1}$.

Proof See Appendix A. □

From the above Theorem, ergodicity of the above process can be established as follows.

Proposition 1 *Under the above setup, it can be obtained that*

$$\lim_{h \to \infty} P(Y_{n+h} = C_i | \mathcal{Y}_n) = p_i,$$

that is, predicted distribution reduces to marginal one if one predicts sufficiently long time ahead.

Proof See Appendix B. □

Although this property was already discussed in Pegram (1980), here we have proved the result using Theorem 2. However, an equivalent result was also provided in Jacobs and Lewis (1978c) for the DAR(p) process. In fact, a generalized result for the NDARMA process is available in Jacobs and Lewis (1983).

3.3 PMA(q) model

Based on the Pegram's operator, Biswas and Song (2009) proposed a stationary MA(q) process, denoted by PMA(q), in the context of discrete time series analysis and is defined as

$$Y_t = (\epsilon_t, \theta_0) * (I(\epsilon_{t-1}), \theta_1) * \cdots * (I(\epsilon_{t-q}), \theta_q),$$

which implies that for every $t \in 0, \pm 1, \pm 2, \ldots$, the conditional probability function takes the form

$$\begin{aligned} P(Y_t = C_i | \epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q}) = \theta_0 I(\epsilon_t = C_i) + \theta_1 I(\epsilon_{t-1} = C_i) \\ + \cdots + \theta_q I(\epsilon_{t-q} = C_i), \end{aligned}$$

where $\theta_i \geq 0$ for all i, and $\sum_{i=0}^{q} \theta_i = 1$. It is easy to see that the marginal distribution of $Y_t \sim D\{(C_i, p_i), i = 0, 1, \ldots, k\}$ for all t. It is to be noted that the PMA(q) process due to Biswas and Song (2009) is indeed equivalent to the DMA(q) model proposed by Jacobs and Lewis (1978a, b).

3.3.1 Coherent forecasting

Consider a stationary PMA(1) model, then the h-step ahead forecasting distribution can be obtained as follows:

For $h = 1$,

$$\begin{aligned} p_1(i) &= P(Y_{n+1} = C_i | \mathcal{Y}_n) \\ &= P(Y_{n+1} = C_i | Y_n) \\ &= \theta_0\theta_1\{I(Y_n = C_i) - p_i\} + p_i, \end{aligned}$$

and for $h > 1$,

$$p_h(i) = P(Y_{n+h} = C_i | Y_n) = p_i.$$

In general, for a stationary PMA(q) model, the h-step ahead forecasting distribution is somewhat complicated with the following representation. For $1 \le h \le q$ and $l = q - 1$,

$$\begin{aligned} p_h(i) &= P(Y_{n+h} = C_i | Y_n = C_{i_0}, \ldots, Y_{n-l} = C_{i_l}) \\ &= \frac{\sum_{r_h=0}^{q}\sum_{r_0=0}^{q}\cdots\sum_{r_l=0}^{q}\theta_{r_h}\theta_{r_0}\cdots\theta_{r_l}P(\epsilon_{n+h-r_h} = C_i, \epsilon_{n-r_0} = C_{i_0}, \ldots, \epsilon_{n-l-r_l} = C_{i_l})}{\sum_{r_0=0}^{q}\cdots\sum_{r_l=0}^{q}\theta_{r_0}\cdots\theta_{r_l}P(\epsilon_{n-r_0} = C_{i_0}, \ldots, \epsilon_{n-l-r_l} = C_{i_l})}, \end{aligned} \tag{3.7}$$

and for $h > q$, $p_h(i) = P(Y_{n+h} = C_i | \mathcal{Y}_n) = p_i$.

An explicit expression of the h-step ahead forecasting distribution for the PMA(2) model is derived in Appendix C.

Thus, the expression for the h-step ahead forecasting distribution of Y_{n+h} given the observed values $Y_1, \ldots, Y_n$ is quite cumbersome for $h \ge 2$. To avoid such complicated results, we suggest to use the following alternative, the h-step ahead forecasting distribution of Y_{n+h} given only the present observed value Y_n, to obtain the h-step ahead coherent forecasting. The advantage of using the following forecasting distribution is that it has a nice and simple expression for all h. Specifically, for $0 < h \le q$, we have

$$\begin{aligned} P(Y_{n+h} = C_i | Y_n = C_j) &= \frac{P(Y_{n+h} = C_i, Y_n = C_j)}{P(Y_n = C_j)} \\ &= \left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\{I(i = j) - p_i\} + p_i, \end{aligned} \tag{3.8}$$

and $P(Y_{n+h} = C_i | Y_n) = p_i$ for $h > q$.

To study the difference between the conditional distribution of Y_{n+h} given Y_n presented in (3.8) and the true forecasting distribution given in (3.7), we carry out one simulation study for the PMA(2) process with different possible choices of the model parameters. We reported the results based on $n = 500$ with model parameters $(\theta_0, \theta_1, \theta_2) = (0.2, 0.6, 0.2)$ and the marginal distribution $\mathbf{p} = (0.2, 0.1, 0.5, 0.15, 0.05)$ defined on the state space $S = \{0, 1, 2, 3, 4\}$. Based on the simulated data, we obtained the exact forecasting distribution using the formula given in Appendix C and the conditional

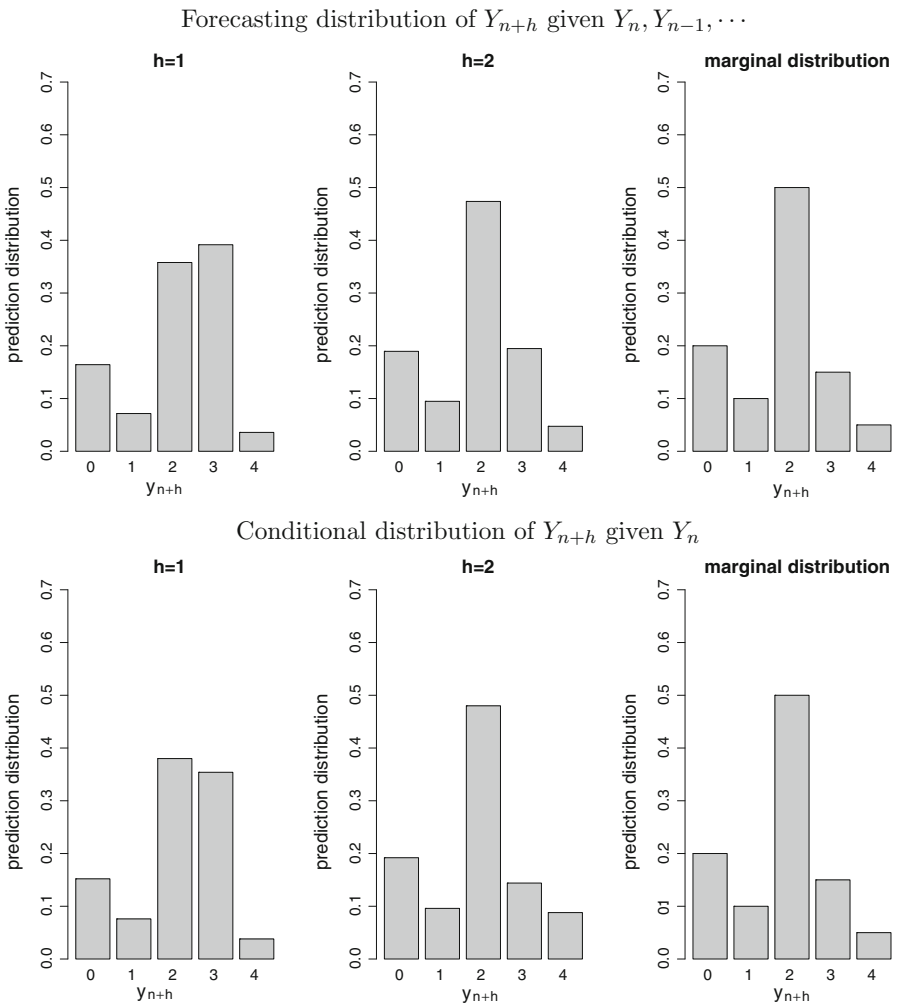

Fig. 1 h-step ahead forecasting and conditional distributions for the PMA(2) process with $(\theta_0, \theta_1, \theta_2) = (0.2, 0.6, 0.2)$ and marginal distribution $\mathbf{p} = (0.2, 0.1, 0.5, 0.15, 0.05)$

distribution of Y_{n+h} given the present observation Y_n given in (3.8). The fitted forecasting distribution and the fitted conditional distribution are presented in Fig. 1. As one can see, no significant difference is visualized. Therefore, one can use the conditional distribution given in Eq. (3.8) as an alternative to the actual forecasting distribution given in Eq. (3.7) whose expression is quite cumbersome to handle while making the coherent forecasting.

3.4 PARMA(p, q) model

Pegram's operator-based ARMA(p, q) model, denoted by PARMA(p,q) due to Biswas and Song (2009) (which is equivalent to NDARMA model by Jacobs and

Lewis 1983), can be constructed by combining the PAR(p) and the PMA(q) models as follows:

$$Y_t = (I(Y_{t-1}), \phi_1) * \cdots * (I(Y_{t-p}), \phi_p) * (\epsilon_t, \theta_0) * (I(\epsilon_{t-1}), \theta_1) * \cdots * (I(\epsilon_{t-q}), \theta_q),$$

which implies that for every $t = 0, \pm 1, \pm 2, \ldots$, the conditional distribution takes the form

$$\begin{aligned} &P(Y_t = C_j | Y_{t-1}, \ldots, Y_{t-p}, \epsilon_t, \ldots, \epsilon_{t-q}) \\ &\quad = \phi_1 I(Y_{t-1} = C_j) + \cdots + \phi_{t-p} I(Y_{t-p} = C_j) \\ &\qquad + \theta_0 I(\epsilon_t = C_j) + \cdots + \theta_q I(\epsilon_{t-q} = C_j), \end{aligned}$$

with $\theta_j \geq 0$ for all j, $\phi_i \geq 0$ for all i, and $\sum_{i=1}^{p} \phi_i + \sum_{j=0}^{q} \theta_j = 1$.

In particular, the PARMA(1,1) model takes the form

$$Y_t = (I(Y_{t-1}), \phi_1) * (\epsilon_t, \theta_0) * (I(\epsilon_{t-1}), \theta_1),$$

with $\phi_1, \theta_0, \theta_1 \geq 0$ and $\phi_1 + \theta_0 + \theta_1 = 1$. Marginal stationarity is guaranteed.

It is easy to obtain the h-step ahead forecasting distribution for the PARMA(1,1) model. For $h = 1$, it is given by

$$P(Y_{n+1} = C_i | Y_n = C_j) = \phi_1 I(j = i) + \theta_0 p_i + \theta_1 \frac{\{\theta_0 I(j = i) + (1 - \theta_0) p_j\} p_i}{p_j},$$

and for $h > 1$,

$$P(Y_{n+h} = C_i | Y_n = C_j) = \phi_1^h I(j = i) + (1 - \phi_1^h) p_i.$$

The forecasting distribution for the PARMA(p,1) model can similarly be obtained as

$$\begin{aligned} p_1(i) &= P(Y_{n+1} = C_i | Y_n = C_{i_0}, \ldots, Y_{n-p+1} = C_{i_{p-1}}) \\ &= \phi_1 I(i_0 = i) + \cdots + \phi_p I(i_{p-1} = i) + \theta_0 p_i \\ &\quad + \theta_1 \frac{\{\theta_0 I(i_0 = i) + (1 - \theta_0) p_{i_0}\} p_i}{p_{i_0}} \\ &= \boldsymbol{\phi}^T \mathbf{e} + \theta_0 p_i + \theta_1 \frac{\{\theta_0 I(i_0 = i) + (1 - \theta_0) p_{i_0}\} p_i}{p_{i_0}}, \end{aligned}$$

where $\mathbf{e} = \big(I(i_0 = i), I(i_1 = i), \ldots, I(i_{p-1} = i)\big)^T$ and for $h > 1$,

$$\begin{aligned} p_h(i) &= \eta_{h1} I(i_0 = i) + \cdots + \eta_{hp} I(i_{p-1} = i) + (1 - \eta_{h1} - \cdots - \eta_{hp}) p_i \\ &= \boldsymbol{\eta}_h^T \mathbf{e} + \left(1 - \boldsymbol{\eta}_h^T \mathbf{1}\right) p_i, \end{aligned}$$

where the h-step ahead parameter $\boldsymbol{\eta}_h$ is given in (3.6). Similarly, for the PARMA(p,2) model and for $h = 1$ we have,

$$\begin{aligned} p_1(i) &= \phi_1 I(i_0 = i) + \cdots + \phi_p I(i_{p-1} = i) + \theta_0 p_i \\ &\quad + \theta_1 \frac{\{\theta_0 I(i_0 = i) + (1-\theta_0) p_{i_0}\} p_i}{p_{i_0}} \\ &\quad + \theta_2 \frac{\{\theta_0 I(i_1 = i) + (1-\theta_0) p_{i_1}\} p_i}{p_{i_1}} \\ &= \boldsymbol{\phi}^T \mathbf{e} + \theta_0 p_i + \theta_1 \frac{\{\theta_0 I(i_0 = i) + (1-\theta_0) p_{i_0}\} p_i}{p_{i_0}} \\ &\quad + \theta_2 \frac{\{\theta_0 I(i_1 = i) + (1-\theta_0) p_{i_1}\} p_i}{p_{i_1}}, \end{aligned}$$

and for $h = 2$,

$$\begin{aligned} p_2(i) &= \phi_1 p_1(i) + \phi_2 I(i_1 = i) + \cdots + \phi_p I(i_{p-1} = i) \\ &\quad + \theta_0 p_i + \theta_1 p_i + \theta_2 \frac{\{\theta_0 I(i_0 = i) + (1-\theta_0) p_{i_0}\} p_i}{p_{i_0}}, \end{aligned}$$

and for $h > 2$,

$$\begin{aligned} p_h(i) &= \eta_{h1} I(i_0 = i) + \cdots + \eta_{hp} I(i_{p-1} = i) + (1 - \eta_{h1} - \cdots - \eta_{hp}) p_i \\ &= \boldsymbol{\eta}_h^T \mathbf{e} + \left(1 - \boldsymbol{\eta}_h^T \mathbf{1}\right) p_i. \end{aligned}$$

It can be further extended for the PARMA(p,q) model.

4 Coherent forecasting for the MTD model

4.1 MTD model

The MTD model was introduced by Raftery (1985) and it bypasses the problem of an exponentially increasing number of free parameters for a Markov chain by specifying the conditional probability of Y_t given the past as a linear combination of contribution from $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$. More precisely, MTD($p$) model assumes that

$$\begin{aligned} P(Y_t = C_i | Y_{t-1} = C_{i_1}, \ldots, Y_{t-p} = C_{i_p}) &= \sum_{j=1}^{p} \lambda_j P(Y_t = C_i | Y_{t-j} = C_{i_j}) \\ &= \sum_{j=1}^{p} \lambda_j q_{i_j i}, \end{aligned} \tag{4.1}$$

where $i, i_1, \ldots, i_p \in \{0, 1, \ldots, k\}$, $q_{i_j i}$s are elements of the $(k+1) \times (k+1)$ transition probability matrix Q and vector of lag parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$ satisfies $\sum_{j=1}^{p} \lambda_j = 1$, $\lambda_j \geq 0$ for all j, so that the right-hand side of (3.1) lies between 0 and 1.

4.2 h-step ahead forecasting distribution

One-step ahead forecasting distribution follows from the model itself, that is

$$p_1(i) = P(Y_{n+1} = C_i | Y_n = C_{i_1}, \ldots, Y_{n-p+1} = C_{i_p}) = \sum_{l=1}^{p} \lambda_l q_{i_l i}, \tag{4.2}$$

Two-step ahead forecasting distribution is given by

$$\begin{aligned} p_2(i) &= P(Y_{n+2} = C_i | Y_n = C_{i_1}, \ldots, Y_{n-p+1} = C_{i_p}) \\ &= \sum_{j_0=0}^{k} P(Y_{n+2} = C_i | Y_{n+1} = C_{j_0}, Y_n = C_{i_1}, \ldots, Y_{n-p+2} = C_{i_{p-1}}) \\ &\quad \times P(Y_{n+1} = C_{i_0} | Y_n = C_{i_1}, \ldots, Y_{n-p+1} = C_{i_p}) \\ &= \sum_{i_0=0}^{k} \sum_{l=1}^{p} \lambda_l q_{i_{l-1} i} \sum_{k=1}^{p} \lambda_k q_{i_k i_0} = \sum_{l=1}^{p} \sum_{k=1}^{p} \lambda_l \lambda_k \left(\sum_{i_0=0}^{k} q_{i_{l-1} i} q_{i_k i_0} \right). \end{aligned} \tag{4.3}$$

Similarly, three-step ahead forecasting distribution is given by

$$\begin{aligned} p_3(i) &= P(Y_{n+3} = C_i | Y_n = C_{i_2}, \ldots, Y_{n-p+1} = C_{i_{p+1}}) \\ &= \sum_{i_0=0}^{k} \sum_{i_1=0}^{k} P(Y_{n+3} = C_i | Y_{n+2} = C_{i_0}, Y_{n+1} = C_{i_1}, \ldots, Y_{n-p+3} = C_{i_{p-1}}) \\ &\quad \times P(Y_{n+2} = C_{i_0} | Y_{n+1} = C_{i_1}, \ldots, Y_{n-p+2} = C_{i_p}) \\ &\quad \times P(Y_{n+1} = C_{i_1} | Y_n = C_{i_2}, \ldots, Y_{n-p+1} = C_{i_{p+1}}) \\ &= \sum_{l=1}^{p} \sum_{k=1}^{p} \sum_{\delta=1}^{p} \lambda_l \lambda_k \lambda_\delta \left(\sum_{i_0=0}^{k} \sum_{i_1=0}^{k} q_{i_{l-1} i} q_{i_k i_0} q_{i_{\delta+1} i_1} \right). \end{aligned}$$

In a similar fashion, we can extend it for any general h. But it is customary to use this forecasting distribution for h less than equal to 4, after that it works like the marginal distribution. Even the forecasting distribution will also become cumbersome.

5 Coherent forecasting for logistic regression model

5.1 Logistic regression model

Some of the inconsistencies associated with standard time series models for count/binary data can be resolved very elegantly and successfully by logistic time series regression (as standard time series models consider simple linear regression on its lag values but logistic regression consider generalized linear regression on its lag values) though stationarity may not be retained here. In the context of categorical time

series analysis, Fokianos and Kedem (2003) applied the same idea to build regression models for categorical time series. Here, we provide a brief description of the multinomial logistic regression model with covariates as its lag values, discuss the estimation of the associated parameters, and then the h-step ahead forecasting distribution and its theoretical confidence interval.

Let $\{Y_t\},\ t = 1, 2, \ldots, N$ be a categorical time series with $(k+1)$ categories. In other words, for each t, the possible values of Y_t are $C_0, C_1, C_2, \ldots, C_k$. As mentioned earlier, 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 the assignment of numbers to categories, it is helpful to note that the t-th observation of any categorical time series regardless of the measurement scale can be expressed by the vector $\mathbf{Y}_t = (Y_{t0}, \ldots, Y_{tq})$ where $q = k - 1$ with elements

$$Y_{tj} = \begin{cases} 1, & \text{if the } j \text{ th category is observed at time } t, \\ 0, & \text{otherwise,} \end{cases} \tag{5.1}$$

for $t = 1, 2, \ldots, N$ and $j = 0, 1, \ldots, q$. Let us denote by $\boldsymbol{\pi}_t = (\pi_{t0}, \pi_{t1}, \ldots, \pi_{tq})$, the vector of conditional probabilities given $\mathcal{F}_{t-1}$, where

$$\pi_{tj} = P(Y_t = C_j | \mathcal{F}_{t-1}), \quad j = 0, 1, \ldots, q$$

for every $t = 1, 2, \ldots, N$. At times, we refer to the π_{tj} as "transition probabilities". Define $Y_{tk} = 1 - \sum_{j=0}^{q} Y_{tj}$ and $\pi_{tk} = 1 - \sum_{j=0}^{q} \pi_{tj}$.

The multinomial logit model defined by Agresti (2002) is given by

$$\pi_{tj}(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}_i^T \mathbf{z}_{t-1})}{1 + \sum_{j=1}^{k} \exp(\boldsymbol{\beta}_j^T \mathbf{z}_{t-1})}, \qquad j = 0, 1, \ldots, q,$$

and

$$\pi_{tk}(\boldsymbol{\beta}) = \frac{1}{1 + \sum_{j=1}^{k} \exp(\boldsymbol{\beta}_j^T \mathbf{z}_{t-1})}.$$

Here $\boldsymbol{\beta}_j,\ j = 0, 1, \ldots, q$ are d-dimensional regression parameters and $\mathbf{z}_{t-1}$ is corresponding d-dimensional vector of stochastic time-dependent covariates independent of j, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \ldots, \boldsymbol{\beta}_q^T)^T$, denotes $(q+1)d$-dimensional vector of parameters. A typical vector of covariates $\mathbf{z}_{t-1} = (1, \mathbf{Y}_{t-1})^T = (1, Y_{(t-1)0}, Y_{(t-1)1}, \ldots, Y_{(t-1)q})^T$ has dimension $d = q + 2$.

To obtain the maximum partial likelihood estimates (MPLE), we maximize the log partial likelihood function which is given by

$$\log PL(\boldsymbol{\beta}) = \sum_{t=1}^{N} \sum_{j=0}^{k} y_{tj} \log \pi_{tj}(\boldsymbol{\beta}), \tag{5.2}$$

and hence

$$\widehat{\boldsymbol{\beta}}_{mple} = \arg\max_{\boldsymbol{\beta}\in\Theta} \log PL(\boldsymbol{\beta}).$$

5.2 Coherent forecasting

To obtain the h-step ahead forecasting for categorical time series for $h > 1$, we can extend the idea given in Fokianos and Kedem (2003). The 1-step ahead predicted response was obtained by the following rule

$$Y_{n+1} = C_i \Leftrightarrow \max_j \pi_{(n+1)j}(\widehat{\boldsymbol{\beta}}) = \pi_{(n+1)i}(\widehat{\boldsymbol{\beta}}).$$

In a recursive way, in the second step we update this predicted observation to the covariates $\mathbf{z}_{n+1}$ and then obtain $\widehat{\pi}_{(n+2)j}$, $j = 0, 1, \ldots, k$ and use the above rule to obtain two-step ahead forecasting, i.e., Y_{n+2} and repeat this process for $h = 3, 4, \ldots$, to obtain the h-step ahead forecasting values. Note that the h-step ahead forecasting distribution is nothing but $p_h(i) = \pi_{(n+h)i}(\boldsymbol{\beta})$, $i = 0, 1, \ldots, k$, which can be used to obtain the forecasting measures KSD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$), ED($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$), and MAD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$) defined in Sect. 2.

5.3 Confidence interval for the h-step ahead forecasting distribution

The h-step ahead forecasting distribution $p_h(i;\boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$. Using delta method, the 95 % confidence interval for $p_h(i;\boldsymbol{\beta})$ is given by $p_h(i;\widehat{\boldsymbol{\beta}}) \mp 1.96\sigma_h(i;\widehat{\boldsymbol{\beta}})$ where

$$\sigma_h^2(i;\boldsymbol{\beta}) = (\bigtriangledown p_h(i;\boldsymbol{\beta}))^T\{G^{-1}(\boldsymbol{\beta})\}(\bigtriangledown p_h(i;\boldsymbol{\beta})) \quad \text{and} \quad \boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_k^T).$$

Fokianos and Kedem (2003) also suggested a consistent estimator for $G(\boldsymbol{\beta})$ given by $\sum_{t=2}^{N} \mathbf{Z}_{t-1}\Sigma_t(\boldsymbol{\beta})\mathbf{Z}_{t-1}^T$, where

$$\mathbf{Z}_{t-1}^{qd\times q} = \begin{pmatrix} \mathbf{z}_{t-1}^{d\times 1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{t-1}^{d\times 1} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{z}_{t-1}^{d\times 1} \end{pmatrix}.$$

6 Simulation study

To study the finite sample behaviors of the proposed forecasting measures, such as PTP, KSD, ED and MAD, and the cardinality of prediction interval defined in Sect. 2, and to facilitate model comparison through the Akaike information criterion (AIC) and

Table 1 Percentage of times AIC and BIC select the correct model where data are generated from **M1, M2, M3**

Sample size	AIC			BIC		
(n)	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
M1						
100	85.5	14.5	0	99.5	0.5	0
300	92.7	7.5	0	100.0	0	0
500	100	0	0	100.0	0	0
1,000	100	0	0	100.0	0	0
5,000	100	0	0	100.0	0	0
M2						
100	0	100	0	10	90	0
300	0	100	0	0	100	0
500	0	100	0	0	100	0
1,000	0	100	0	0	100	0
5,000	0	100	0	0	100	0
M3						
100	0	0	100	0	0	100
300	0	0	100	0	0	100
500	0	0	100	0	0	100
1,000	0	0	100	0	0	100
5,000	0	0	100	0	0	100

Bayesian information criterion (BIC) and the above forecasting measures, we carried out some simulation studies based on the samples generated from the following three categorical time series models with 4 categories $\{C_0, C_1, C_2, C_3\}$.

(**M1**) PAR(1) model with $\phi = 0.8$ and $\mathbf{p} = (0.2, 0.2, 0.5, 0.1)$,
(**M2**) MTD(1) model with the transition probability matrix

$$\mathbf{Q} = \begin{pmatrix} 0.85 & 0.01 & 0.05 & 0.09 \\ 0.25 & 0.20 & 0.35 & 0.20 \\ 0.05 & 0.10 & 0.80 & 0.05 \\ 0.05 & 0.05 & 0.20 & 0.70 \end{pmatrix}, \quad \text{and}$$

(**M3**) Logistic regression model of order 1 with covariates $\mathbf{z}_{t-1} = (1, \mathbf{Y}_{t-1})^T = (1, Y_{(t-1)1}, Y_{(t-1)2}, Y_{(t-1)3})^T$ and $\boldsymbol{\beta}_0 = (6.80, 5.00, 3.30, 3.90)^T$, $\boldsymbol{\beta}_1 = (2.45, 4.80, 4.05, 3.90)^T$, $\boldsymbol{\beta}_2 = (4.05, 5.35, 6.25, 5.50)^T$.

To begin with, we generated samples of different sizes from the above three cases, namely **M1**, **M2** and **M3** and presented the results in Table 1. Five sample sizes are explored: samples of sizes 100 and 300 are used to study the small sample properties, samples of sizes 500 and 1,000 are used to get an idea about the moderate sample properties, and samples of size 5,000 are used to study the large sample properties.

For a fixed sample size n, we repeated the process 1,000 times and observed the percentage of times AIC and BIC select a particular model from the three models under comparison. Table 1 summarizes the results based on the data generated from the **M1, M2** and **M3**. As expected, most of the times almost in all the cases AIC and BIC selected the true data-generating model, except for the second case **M2**. In case of **M2**, for small sample size (100), BIC selected the PAR(1) model 10 % times as the true model, although the true data-generating mechanism was MTD(1). This is because the MTD model suffers from the large number of parameters which is considered as penalty in BIC.

In the second study, samples of size 150 were generated from all the three cases **M1**, **M2** and **M3**. Then for each cases, we fitted all the three models under comparison, and obtained the forecasting measures—PTP, KSD, ED and MAD for varying h. The results based on 5,000 replications are reported in Table 2. As we can see from the Table 2, for all the three cases, the measures KSD, ED and MAD increase as h increases. It means that forecasting accuracy decreases as one goes far ahead from the present as far as the KSD, ED and MAD are concerned, which is expected. On the other hand, as expected for all the cases, PTP measure decreases as h increases (see Table 2). Another important observation reveals that when the data were generated from **M1**, PAR(1) outperformed others with respect to all the four forecasting measures, whereas MTD(1) and Logistic(1) outperformed when data were generated from **M2** and **M3**, respectively, which is also an expected scenario. Therefore,we may say that in all these cases the above forecasting measures played a significant role in detecting the true model.

In an other study, we repeated the previous exercise, where we simulated samples of size 150 from all the three cases **M1**, **M2** and **M3** to study the forecasting accuracy using the HPP set ($\mathcal{S}_h$). For each data-generating mechanism, we obtained the $100(1-\alpha)$ % HPP set ($\mathcal{S}_h$) for $h = 1, \ldots, 6$ using the true data-generating models which are PAR(1), MTD(1) and Logistic(1) where $\alpha = 0.2$. The results based on all the three data-generating models are presented in Table 3. As we can see, for all the three cases cardinality of $\mathcal{S}_h$ increases as h increases, which implies that to capture the same percentage of true observations as one goes far ahead from the present, one needs a larger HPP set. Therefore, the HPP set would also be a sensible measure to study the forecasting accuracy in the discrete time series analysis especially for categorical time series as far as the interval forecasting is concerned.

7 Real data example: infant sleep status data

Stoffer et al. (1988) reported a collection of 24 categorical time series of infant sleep status which is divided into two groups of 12 each based on their mothers' drinking habit during pregnancy (one group of mothers abstained from drinking alcohol throughout their pregnancy, and the other group used alcohol moderately and consistently throughout their pregnancy), in an EEG study. Each of these 24 time series is observed for 128 min. In this section, we consider one such single time series from the first group.

Table 2 Values of forecasting measures PTP, KSD, ED and MAD for varying h where the data-generating model are **M1, M2, M3**

h-step	PTP(h)			KSD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)			ED($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)			MAD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)		
	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
M1												
1	83.68	83.36	83.62	0.0185	0.0257	0.0247	0.0294	0.0419	0.0418	0.0221	0.0329	0.0349
2	72.26	71.44	70.76	0.0302	0.0408	0.0468	0.0491	0.0671	0.0721	0.0370	0.0529	0.0609
3	62.55	61.22	60.78	0.0376	0.0487	0.0507	0.0622	0.0812	0.0791	0.0473	0.0638	0.0628
4	53.50	52.80	53.00	0.0425	0.0522	0.0512	0.0708	0.0880	0.0879	0.0542	0.0685	0.0725
5	48.68	46.23	45.83	0.0456	0.0524	0.0527	0.0765	0.0903	0.0932	0.0586	0.0693	0.0793
6	42.52	41.14	39.77	0.0489	0.0516	0.0546	0.0803	0.0902	0.0901	0.0616	0.0686	0.0826
M2												
1	84.79	85.90	84.86	0.0671	0.0143	0.0156	0.0780	0.0198	0.0223	0.0586	0.0154	0.0171
2	73.73	74.84	73.93	0.1039	0.0245	0.0246	0.1268	0.0327	0.0356	0.0961	0.0255	0.0278
3	65.65	67.90	65.67	0.1296	0.0325	0.0339	0.1596	0.0426	0.0495	0.1212	0.0334	0.0383
4	59.39	60.53	59.27	0.1455	0.0391	0.0427	0.1817	0.0506	0.0627	0.1381	0.0398	0.0479
5	54.50	55.71	54.59	0.1548	0.0446	0.0502	0.1959	0.0572	0.0730	0.1493	0.0449	0.0559
6	50.64	51.62	51.54	0.1594	0.0491	0.0566	0.2042	0.0627	0.0814	0.1559	0.0494	0.0624
M3												
1	92.49	92.25	93.35	0.0631	0.0152	0.0130	0.0737	0.0205	0.0166	0.0586	0.0159	0.0131
2	86.37	86.59	87.46	0.1003	0.0238	0.0219	0.1176	0.0288	0.0262	0.0923	0.0227	0.0213
3	80.94	81.17	82.89	0.1219	0.0360	0.0289	0.1450	0.0358	0.0338	0.1111	0.0289	0.0277
4	76.14	76.48	77.09	0.1351	0.0361	0.0345	0.1626	0.0419	0.0401	0.1226	0.0342	0.0331
5	72.04	72.08	72.67	0.1429	0.0406	0.0392	0.1738	0.0472	0.0455	0.1302	0.0386	0.0378
6	68.11	67.98	68.81	0.1470	0.0443	0.0430	0.1804	0.0516	0.0501	0.1350	0.0423	0.0416

Table 3 $100(1-\alpha)\,\%$ HPP set of Y_{n+h} given Y_n for varying h with cardinality of the set, where data are generated from **M1, M2, M3**, where $\alpha = 0.2$

h-step	**M1**				**M2**				**M3**			
	$Y_n = 0$		$Y_n = 1$		$Y_n = 0$		$Y_n = 2$		$Y_n = 0$		$Y_n = 2$	
	$\mathcal{S}_h$	$n(\mathcal{S}_h)$	$\mathcal{S}_h$	$n(\mathcal{S}_h)$	$\mathcal{S}_h$	$n(\mathcal{S}_h)$	$\mathcal{S}_h$	$n(\mathcal{S}_h)$	$\mathcal{S}_h$	$n(\mathcal{S}_h)$	$\mathcal{S}_h$	$n(\mathcal{S}_h)$
1	$\{C_0\}$	1	$\{C_1\}$	1	$\{C_0\}$	1	$\{C_2, C_3\}$	2	$\{C_0\}$	1	$\{C_2, C_3\}$	2
2	$\{C_0, C_2\}$	2	$\{C_1, C_2\}$	2	$\{C_0, C_3\}$	2	$\{C_2, C_3\}$	2	$\{C_0, C_2\}$	2	$\{C_2, C_3\}$	2
3	$\{C_0, C_2\}$	2	$\{C_1, C_2\}$	2	$\{C_0, C_3\}$	2	$\{C_0, C_2, C_3\}$	3	$\{C_0, C_2\}$	2	$\{C_0, C_2, C_3\}$	2
4	$\{C_0, C_2\}$	2	$\{C_0, C_1, C_2\}$	3	$\{C_0, C_2\}$	2	$\{C_0, C_2, C_3\}$	3	$\{C_0, C_2\}$	2	$\{C_0, C_2, C_3\}$	3
5	$\{C_0, C_2\}$	2	$\{C_0, C_1, C_2\}$	3	$\{C_0, C_2, C_3\}$	3	$\{C_0, C_2, C_3\}$	3	$\{C_0, C_2\}$	2	$\{C_0, C_2, C_3\}$	3
6	$\{C_0, C_2, C_3\}$	3	$\{C_0, C_1, C_2\}$	3	$\{C_0, C_2, C_3\}$	3	$\{C_0, C_2, C_3\}$	3	$\{C_0, C_2\}$	2	$\{C_0, C_2, C_3\}$	3

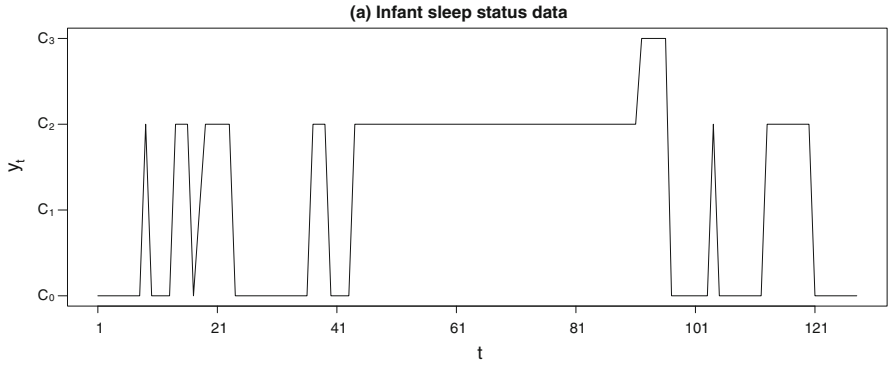

Fig. 2 Plot of the infant sleep status data after combining some states

During minute t, the infant's sleep status was recorded in six categories, namely "qt" being 'quiet sleep' with trace alternate, "qh" being 'quiet sleep' with high voltage, "tr" being 'transitional sleep', "al" being 'active sleep' with low voltage, "ah" being 'active sleep' with high voltage, and "aw" being 'awake'. Note that the number of parameters to be estimated is 6 for PAR(1) model and is 30 for MTD(1) model which is quite large against the data size 128. On the other hand, since number of categories is 6, if we want to fit the logistic regression model, $\mathbf{Y}_t$ has 5 components, i.e., $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, Y_{t3}, Y_{t4}, Y_{t5})$. Therefore, the number of parameters to be estimated to fit the logistic regression model of order 1 with covariates $\mathbf{z}_{t-1} = (1, \mathbf{Y}_{t-1}) = (1, Y_{(t-1)1}, Y_{(t-1)2}, Y_{(t-1)3}, Y_{(t-1)4}, Y_{(t-1)5})$ will be 30, and it is done using the partial likelihood method given in Eq. (5.2). Note that partial likelihood estimates of 30 parameters based on the data of size 128 may not be so reliable. Therefore, to bypass the problem, we reduced the number of categories from 6 to 4 by combining the quite states and active states as suggested in Stoffer et al. (2000). Hence the numbers of parameters to be estimated for the PAR(1) model becomes 4, and it is 12 for both the MTD(1) and Logistic(1) models. After combining the quite states and active states, the new labels of the categories are given by

$$\text{qt} \equiv C_0, \quad \text{qh} \equiv C_0, \quad \text{tr} \equiv C_1, \quad \text{al} \equiv C_2, \quad \text{ah} \equiv C_2, \quad \text{aw} \equiv C_3. \tag{7.1}$$

The proportion of times spent by an infant in the combined sleep status C_0, C_1, C_2 and C_3 given in (7.1) is 0.414, 0.008, 0.539, and 0.039, respectively. This indicates that the infant spent maximum time in the active sleep. The combined data are plotted in Fig. 2.

It is important to mention that although the infant sleep status data are of ordinal in nature, it may not be appropriate to use the ACF and PACF plots to choose the correct order. This is because the values of ACF and PACF depend on the actual numerical scaling of the categories and it changes from one scaling to another scaling of the categories. In practice, there does not exist any unique numerical scaling for such ordinal categories. We may at most say that the four scale values should be $C_0 < C_1 < C_2 < C_3$, and cannot specify the values of C_0, C_1, C_2, C_3. Hence some

Table 4 Estimated values of $\kappa(h)$, $v(h)$, $A_v^{(\tau)}(h)$ and Cohen's κ-based partial autocorrelation ($\rho_{\mathrm{p}}(h)$) for the infant sleep status data

Lag h	$\widehat{\kappa}(h)$	$\widehat{v}(h)$	$\sqrt{\widehat{A}_v^{(\tau)}(h)}$	$\widehat{\rho}_{\mathrm{p}}(h)$
1	0.7651	0.6489	0.6085	**0.7651**
2	0.6152	0.5092	0.4019	0.0720
3	0.4779	0.3842	0.2594	−0.0339
4	0.3985	0.3115	0.2082	0.0603
5	0.3633	0.3114	0.2167	0.0891
6	0.3123	0.2924	0.1843	−0.0236

alternative measures of serial association, which do not depend on the numerical scaling of the categories, should be used to select the order of the process. Weiß and Göb (2008) established one Theorem for an empirical justification of the adequacy of the NDARMA(p, q) model to the observed categorical data (see Theorem 5.2 in Weiß and Göb 2008). The Theorem says that the estimates $\widehat{\kappa}(h)$ for Cohen's κ, $\widehat{v}(h)$ for Cramer's v, and the square root of estimate $\widehat{A}_v^{(\tau)}(h)$ for Goodman and Kruskal's τ of different lag values h will be approximately equal if the NDARMA(p, q) is adequate to the data. The formulae of these measures and their estimates are given in detail in Weiß and Göb (2008), Weiß (2011, 2013). Then to select the order of the NDARMA(p, q) model, they proposed to observe the usual PACF, $\rho_{\mathrm{p}}(h)$ based on the estimates $\widehat{\kappa}(h)$ for Cohen's κ in place of the ACF $\rho(h)$.

For the infant sleep status data, we obtained the values of these measures for various lag values and present the results in Table 4. As we can see, all the three measures, namely Cohen's κ, Cramer's v and square root of Goodman and Kruskal's τ are close enough to fit the data by a PAR(p) process. On the other hand, Cohen's κ estimates based estimates of PACFs, $\widehat{\rho}_{\mathrm{p}}(h)$, are about 0 for $h > 1$. Therefore, a PAR(1) model, which is same as DAR(1) model, will be an appropriate fit to the data.

In addition, to study the effectiveness of the Cohen's κ measure, we derived it for the PAR(1) model which came out to be ϕ^h and hence it decreases as the lag value h increases. Based on this result, we performed one simulation study. We generated samples of sizes $n = 200, 1,000, 10,000$ from the PAR(1) model with number of categories 4, for mixing parameter $\phi = 0.4, 0.6, 0.8$ and common marginal distribution $\mathbf{p} = (0.414, 0.008, 0.539, 0.039)$. Figure 3 displays the values of theoretical $\kappa(h)$ (which is colored in black) with the empirical $\kappa(h)$ (which is colored in gray) for varying h. We see that as the sample size increases the empirical $\kappa(h)$ coincides with the theoretical $\kappa(h)$. Based on this observation, we fitted the PAR(1) model to the infant sleep status data and obtained the empirical and theoretical values of $\kappa(h)$ for various values of h and presented it in Fig. 4. As we can see from Fig. 4, the empirical $\kappa(h)$ obtained from the data coincides with the fitted PAR(1) model.

The transition probabilities for MTD(1) model are obtained through sample proportions, whereas the parameters for PAR(1) and logistic regression models of order 1 are estimated using partial likelihood method. The estimated value of the mixing parameter ϕ of the PAR(1) model is 0.78, which indicates that a large number of paired observations (Y_t, Y_{t-1}) with $Y_t = Y_{t-1}$ is present in the data. Therefore, the PAR(1) model is a competing alternative to the data. The other parameter associated with the PAR(1) model is the marginal distribution $\mathbf{p}$ which is estimated

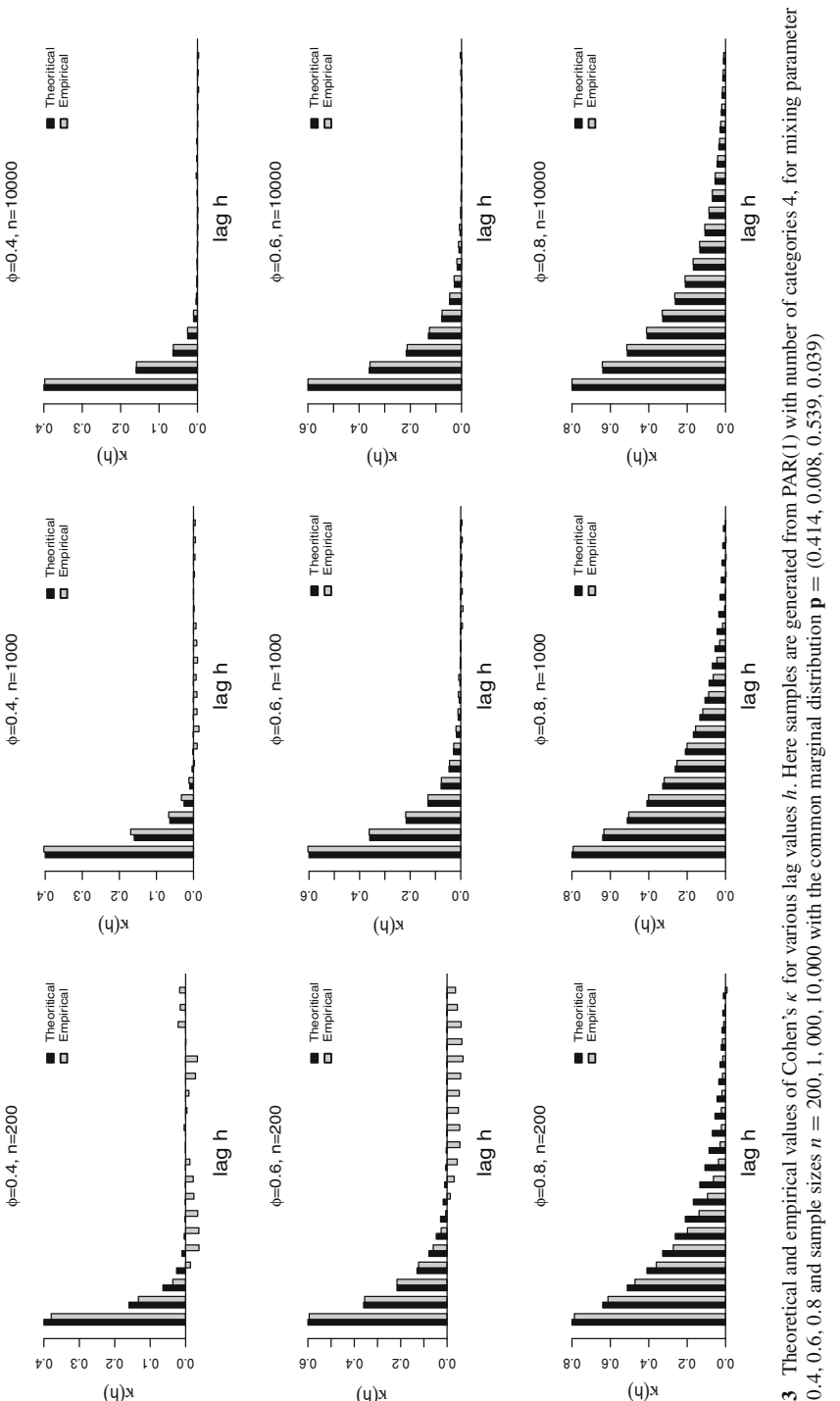

Fig. 3 Theoretical and empirical values of Cohen's κ for various lag values h. Here samples are generated from PAR(1) with number of categories 4, for mixing parameter $\phi = 0.4, 0.6, 0.8$ and sample sizes $n = 200, 1,000, 10,000$ with the common marginal distribution $\mathbf{p} = (0.414, 0.008, 0.539, 0.039)$

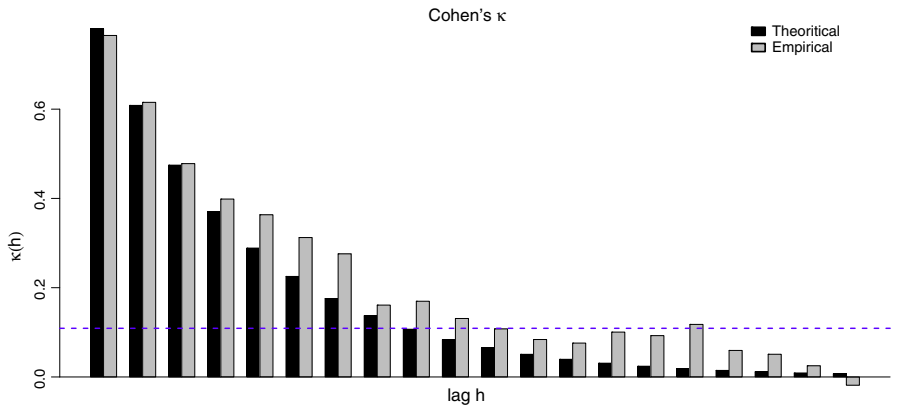

Fig. 4 Plot of Cohen's κ for varying lag values for the infant sleep data

as (0.414, 0.008, 0.539, 0.0391). Similarly the transition probability matrix (tpm) $\mathbf{Q}$ associated with MTD(1) model is estimated as

$$\begin{pmatrix} 0.869 & 0.019 & 0.115 & 0 \\ 0 & 0 & 1 & 0 \\ 0.087 & 0 & 0.898 & 0.014 \\ 0.200 & 0 & 0 & 0.800 \end{pmatrix}.$$

To fit the logistic regression model, we used the setup discussed in Eq. (5.1) in Sect. 5. Note that, after combining the states the data has four categories and hence $\mathbf{Y}_t$ has three components, i.e., $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, Y_{t3})^T$. Based on this multivariate representation, we plotted sample autocorrelation and cross-correlation in Fig. 5. As one can see, there is a decreasing pattern in the first and last plots in Fig. 5, which indicates that Y_t only depends on its lagged values, and there is no periodical term (e.g., sinusoidal term) in its covariates. Therefore, we fitted logistic regression model with covariates $\mathbf{z}_{t-1} = (1, \mathbf{Y}_{t-1})^T = (1, Y_{(t-1)1}, Y_{(t-1)2}, Y_{(t-1)3})^T$ (we called it Logistic(1) model with intercept). The parameters associated with the model were estimated as

$$\boldsymbol{\beta_0} = (6.80, 5.00, 3.30, 3.90)^T, \quad \boldsymbol{\beta_1} = (2.45, 4.80, 4.05, 3.90)^T, \quad \text{and}$$
$$\boldsymbol{\beta_2} = (4.05, 5.35, 6.25, 5.50)^T.$$

After fitting the above models, we obtained the AIC and BIC for all the three models and presented it in Table 5. As we can see, PAR(1) model has the lowest AIC and BIC values. In addition, we obtained the PTP measure by dividing the data into two parts. First part, the training part consisting first 110 observations, was used to fit the models under comparison, and we obtained the single PTP measure based on the remaining 18 observations, which is presented in Table 5. As we can see, the PAR(1) outperforms MTD(1) and Logistic(1) in terms of predicting the true observations. Hence, overall the PAR(1) model fitted the data best among these three competing models.

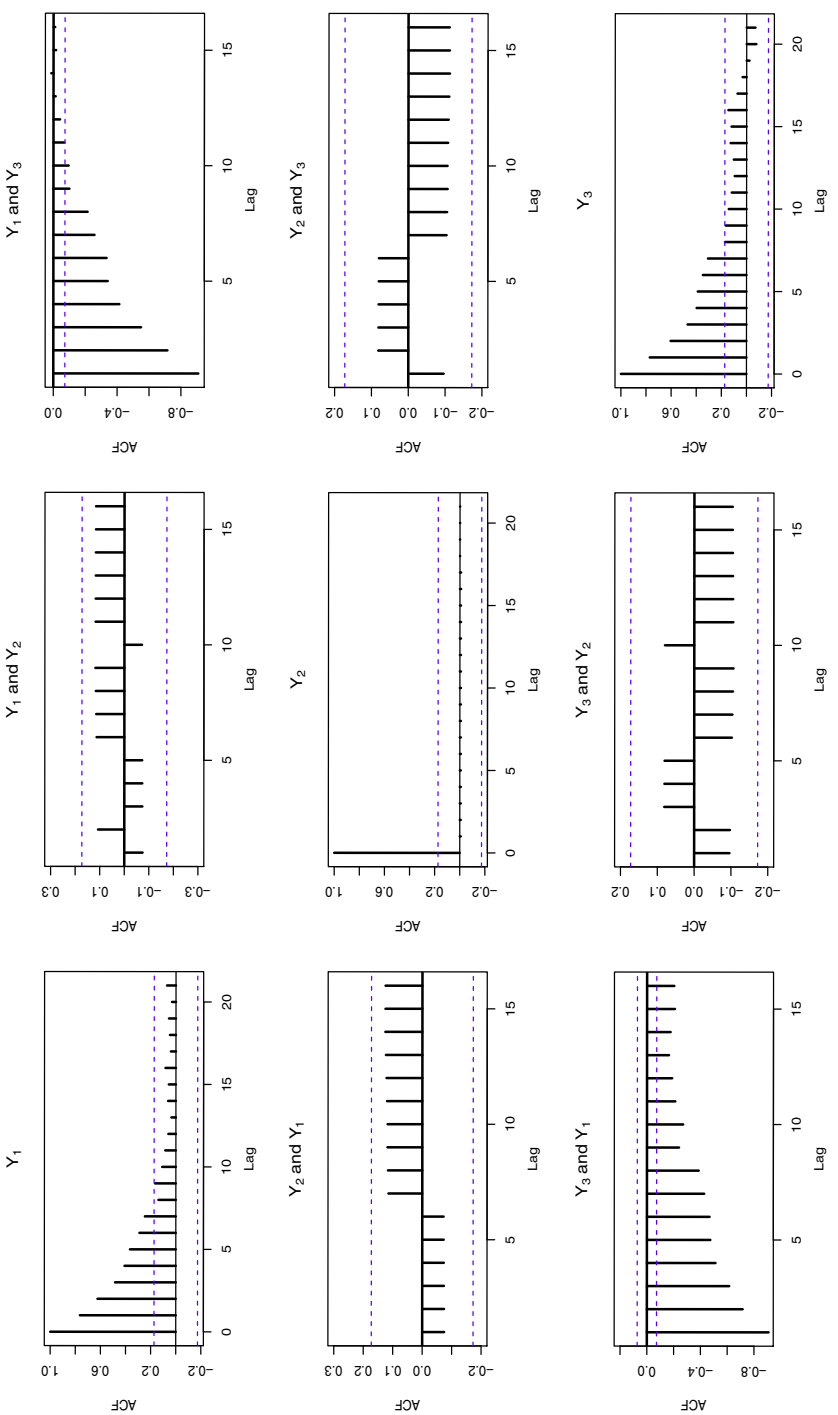

Fig. 5 Sample autocorrelation and cross-correlation functions for the infant sleep status data

Table 5 Infant sleep status data analysis

Model	AIC	BIC	PTP
PAR(1)	**122.94**	**134.35**	**33.33**
MTD(1)	126.87	161.10	27.78
Logistic(1)	134.88	180.51	27.78

8 Concluding remarks

The basic objective of the present paper is to study the different methods of coherent forecasting and their forecasting accuracy based on some forecasting measures, which has been defined in Sect. 2 including forecasting interval in the context of time series of discrete data, especially for categorical data. Theoretical results and some simulation studies with a real data analysis on infant sleep status have illustrated the proposed methods.

Note that when the time series data are categorical, popular measures for studying forecasting accuracy like PRMSE and PMAE cannot be used. Therefore, to study the forecasting accuracy for categorical time series, here we have defined different measures, namely PTP, KSD, ED and MAD. Through some extensive simulation studies, efficacy of these measures have been checked. In addition, we have introduced a different notion of interval forecasting for categorical time series analysis whose efficacy has also been checked using some simulation results. Hence, we can say that these measures can be used in practice for the analysis of categorical time series data.

On the other side, a comparison study has been performed using those forecasting methods. Note that, Pegram's operator-based AR(p), MA(q) or ARMA(p,q) models are applicable for both count and categorical data (see, e.g., Biswas and Song 2009; Biswas and Guha 2009). However, the MTD model due to Raftery (1985) and the logistic regression model due to Fokianos and Kedem (2003) have a serious drawback that the number of parameters to be estimated is very large for large number (greater than 3) of categories which makes it difficult to implement. In addition, as observed in the simulation study, even though the data are generated from the MTD model, the BIC may be larger than the Pegram's AR model due to the large number of parameters in the MTD model. As a result, the BIC may select some other competing model as the true model even though the data-generating mechanism is MTD model. On the other hand, the logistic regression models lack stationarity unless the parameters are appropriately adjusted. The Pegram's ARMA model is very simple-minded and it is stationary and involves smaller number of parameters than the MTD and the logistic models. Also it has many elegant theoretical properties. Hence, it can be a good choice in many practical situations.

Acknowledgments The authors wish to thank the three anonymous referees and the associate editor for their careful reading and constructive suggestions which led to this improved version of the paper.

Appendix

Appendix A : Proof of Theorem 2

From the model (3.2), the 1-step ahead conditional distribution is given by

$$\begin{aligned} p_1(i|i_1,\ldots,i_p) &= P(Y_{n+1}=C_i|Y_n=C_{i_1},\ldots,Y_{n-p+1}=C_{i_p}) \\ &= \eta_{11}I(i_1=i)+\cdots+\eta_{1p}I(i_p=i)+(1-\eta_{11}-\cdots-\eta_{1p})p_i, \end{aligned}$$

with $\eta_{1l}=\phi_l, l=1,\ldots,p$. Then the two-step ahead conditional distribution is given by

$$\begin{aligned} p_2(i|i_1,\ldots,i_p) &= P(Y_{n+2}=C_i|Y_n=C_{i_1},\ldots,Y_{n-p+1}=C_{i_p}) \\ &= \sum_{j=0}^{k} P(Y_{n+2}=C_i|Y_{n+1}=C_j,Y_n=C_{i_1},\ldots,Y_{n-p+2}=C_{i_{p-1}}) \\ &\quad \times P(Y_{n+1}=C_j|Y_n=C_{i_1},\ldots,Y_{n-p+1}=C_{i_p}) \\ &= \sum_{j=0}^{k}\left\{\eta_{11}I(j=i)+\cdots+\eta_{1p}I(i_{p-1}=i)+(1-\eta_{11}-\cdots-\eta_{1p})p_i\right\} \\ &\quad \times\left\{\phi_1 I(i_1=j)+\cdots+\phi_p I(i_p=j)+(1-\phi_1-\cdots-\phi_p)p_j\right\} \\ &= \eta_{21}I(i_1=i)+\cdots+\eta_{2p}I(i_p=i)+(1-\eta_{21}-\cdots-\eta_{2p})p_i \end{aligned}$$

where $\boldsymbol{\eta_2}=\boldsymbol{\Phi\phi}$. So the result is true for $h=2$. Let it be true for $(h-1)$, that is $\boldsymbol{\eta_{h-1}}=\boldsymbol{\Phi}^{h-2}\boldsymbol{\phi}$. Then by induction it is straightforward to show that the h-step ahead conditional distribution is given by (3.5).

Appendix B : Proof of Theorem 3

To prove the Theorem 3, it is enough to show that $\lim_{h\to\infty}\eta_{hi}=0$ for all i. To show this we use the result that for any $n\times n$ matrix A with its eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_s$, $\lim_{k\to\infty}A^k=0$ if the spectral radius of A, $\rho(A)<1$ where $\rho(A)=\max\{|\lambda_1|,|\lambda_2|,\ldots,|\lambda_s|\}$ (See Atkinson 2008). Outline of the proof is given follows.

From the Jordan normal theorem, for any $n\times n$ matrix A, there exist a non-singular matrix V and a block diagonal matrix J such that

$$A=VJV^{-1}$$

for

$$J=\begin{pmatrix} J_{m_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{m_2}(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{m_s}(\lambda_s) \end{pmatrix},$$

where the $m_i\times m_i$ matrix $J_{m_i}(\lambda_i)$ being

$$J_{m_i}(\lambda_i)=\begin{pmatrix} \lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{pmatrix}.$$

Now

$$A^k = V J^k V^{-1}$$

and, since J is block diagonal,

$$J^k = \begin{pmatrix} J_{m_1}^k(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{m_2}^k(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{m_s}^k(\lambda_s) \end{pmatrix}.$$

Now a standard result on the kth power of an $m \times m$ Jordan block states that, for $k \geq m$,

$$J_m^k(\lambda) = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \cdots & \binom{k}{m-1}\lambda^{k-m+1} \\ 0 & \lambda^k & \binom{k}{1}\lambda^{k-1} & \cdots & \binom{k}{m-2}\lambda^{k-m+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda^k \end{pmatrix}.$$

Since $\rho(A) < 1$, i.e., $|\lambda_i| < 1$ for all i and $\lim_{k\to\infty} \binom{k}{i}\lambda^{k-i} = 0$, and hence $\lim_{k\to\infty} J_m^k(\lambda) = 0$, . This implies that $\lim_{k\to\infty} J^k = 0$. Therefore,

$$\lim_{k\to\infty} A^k = \lim_{k\to\infty} V J^k V^{-1} = V(\lim_{k\to\infty} J^k)V^{-1} = 0.$$

Note that the eigenvalues of $\mathbf{\Phi}$ are $\phi_1, \ldots, \phi_p$ all of which lie between 0 and 1, and hence $\lim_{h\to\infty} \mathbf{\Phi}^h = 0$. Consequently $\lim_{h\to\infty} \boldsymbol{\eta}_h = \lim_{h\to\infty} \mathbf{\Phi}^{h-1}\boldsymbol{\phi} = (\lim_{h\to\infty} \mathbf{\Phi}^{h-1})\boldsymbol{\phi} = 0$.

Appendix C: Pegram's MA(2) model

Here for $h = 1$,

$$P(Y_{n+1} = C_i | Y_n = C_j, Y_{n-1} = C_k) = \frac{\sum_{r=0}^{2}\sum_{s=0}^{2}\sum_{t=0}^{2} \theta_r\theta_s\theta_t P(\epsilon_{n+1-r} = C_i, \epsilon_{n-s} = C_j, \epsilon_{n-1-t} = C_k)}{\sum_{s=0}^{2}\sum_{t=0}^{2} \theta_s\theta_t P(\epsilon_{n-s} = C_j, \epsilon_{n-1-t} = C_k)},$$

where

$$\begin{aligned}P(\epsilon_{n+1-r}=C_i,\epsilon_{n-s}=C_j,\epsilon_{n-1-t}=C_k) &= p_ip_jp_kI(r-1\neq s\neq t+1)\\ &+ p_ip_jI(j=k)I(r-1\neq s=t+1)\\ &+ p_ip_jI(i=k)I(r-1=t+1\neq s)\\ &+ p_ip_kI(i=j)I(r-1=s\neq t+1)\\ &+ p_iI(i=j=k)I(r-1=s=t+1)\end{aligned}$$

and

$$P(\epsilon_{n-s}=C_j,\epsilon_{n-1-t}=C_k)=p_jp_kI(s\neq t+1)+p_jI(j=k)I(s=t+1).$$

Similarly for $h=2$,

$$\begin{aligned}&P(Y_{n+2}=C_i|Y_n,\ldots,Y_1)\\ &=P(Y_{n+2}=C_i|Y_n=C_j,Y_{n-1}=C_k)\\ &=\frac{P(Y_{n+2}=C_i,Y_n=C_j,Y_{n-1}=C_k)}{P(Y_n=C_j,Y_{n-1}=C_k)}\\ &=\frac{\sum_{r=0}^{2}\sum_{s=0}^{2}\sum_{t=0}^{2}\theta_r\theta_s\theta_tP(\epsilon_{n+2-r}=C_i,\epsilon_{n-s}=C_j,\epsilon_{n-1-t}=C_k)}{\sum_{s=0}^{2}\sum_{t=0}^{2}\theta_s\theta_tP(\epsilon_{n-s}=C_j,\epsilon_{n-1-t}=C_k)}\end{aligned}$$

where

$$\begin{aligned}P(\epsilon_{n+2-r}=C_i,\epsilon_{n-s}=C_j,\epsilon_{n-1-t}=C_k) &= p_ip_jp_kI(r-2\neq s\neq t+1)\\ &+ p_ip_jI(j=k)I(r-2\neq s=t+1)\\ &+ p_ip_jI(i=k)I(r-2=t+1\neq s)\\ &+ p_ip_kI(i=j)I(r-2=s\neq t+1)\\ &+ p_iI(i=j=k)I(r-2=s=t+1),\end{aligned}$$

and

$$P(\epsilon_{n-s}=C_j,\epsilon_{n-1-t}=C_k)=p_jp_kI(s\neq t+1)+p_jI(j=k)I(s=t+1).$$

And for $h>2$, $P(Y_{n+h}=C_i|Y_n,Y_{n-1},\ldots)=p_i$.

References

Agresti, A.: Categorical data analysis. Wiley, New Jersey (2002)

Al-Osh, M., Alzaid, A.A.: First-order integer-valued autoregressive (INAR(1)) process. J. Time Ser. Anal. **8**(3), 261–275 (1987)

Alzaid, A., Al-Osh, M.: An integer-valued pth-order autoregressive structure (INAR(p)) process. J. Appl. Probab. **27**, 314–324 (1990)

Atkinson, K.: An introduction to numerical analysis. Wiley, New Delhi (2008)
Berchtold, A., Raftery, A.E.: The mixture transition distribution model for high-order Markov chains and non-gaussian time series. Stat. Sci. **17**(3), 328–356 (2002)
Biswas, A., Guha, A.: Time series analysis of categorical data using auto-mutual information. J. Stat. Plan. Inference **139**(9), 3076–3087 (2009)
Biswas, A., Song, P.X.-K.: Discrete-valued ARMA processes. Stat. Prob. Lett. **79**(17), 1884–1889 (2009)
Brockwell, P.J., Davis, R.A.: Time series: theory and methods. Springer, Berlin (2002)
Bu, R., McCabe, B.: Model selection, estimation and forecasting in INAR(p) models: a likelihood-based Markov chain approach. Int. J. Forecast. **24**(1), 151–162 (2008)
Carruth, J., Tygert, M., Ward, R.: A comparison of the discrete Kolmogorov–Smirnov statistic and the euclidean distance (2012). arXiv:1206.6367
Fokianos, K., Kedem, B.: Regression theory for categorical time series. Stat. Sci. **18**(3), 357–376 (2003)
Freeland, R.K., McCabe, B.P.: Forecasting discrete valued low count time series. Int. J. Forecast. **20**(3), 427–434 (2004)
Jacobs, P.A., Lewis, P.A.: Discrete time series generated by mixtures III: Autoregressive processes (DAR(p)). Technical report, Naval Postgraduate School, Monterey (1978c)
Jacobs, P.A., Lewis, P.A.: Discrete time series generated by mixtures. I: correlational and runs properties. J. Royal Stat. Soc. Ser. B (Methodological) **40**(1), 94–105 (1978a)
Jacobs, P.A., Lewis, P.A.: Discrete time series generated by mixtures II: asymptotic properties. J. Royal Stat. Soc. Ser. B (Methodological) **40**(2), 222–228 (1978b)
Jacobs, P.A., Lewis, P.A.: Stationary discrete autoregressive–moving average time series generated by mixtures. J. Time Ser. Anal. **4**(1), 19–36 (1983)
Jung, R.C., Tremayne, A.R.: Coherent forecasting in integer time series models. Int. J. Forecast. **22**(2), 223–238 (2006)
McKenzie, E.: Discrete variate time series. Handbook of statistics 21, 573–606 (2003)
McKenzie, E.: Some simple models for discrete variate time series. J. Am. Water Resour. Assoc. **21**(4), 645–650 (1985)
McKenzie, E.: Some ARMA models for dependent sequences of Poisson counts. Adv. Appl. Probab. **20**(4), 822–835 (1988)
Pegram, G.: An autoregressive model for multilag Markov chains. J. Appl. Probab. **17**(2), 350–362 (1980)
Raftery, A.E.: A model for high-order Markov chains. J. Royal Stat. Soc. Ser. B (Methodological) **47**(3), 528–539 (1985)
Silva, N., Pereira, I., Silva, M.E.: Forecasting in INAR(1) model. REVSTAT Stat. J. **7**(1), 119–134 (2009)
Stoffer, D.S., Scher, M.S., Richardson, G.A., Day, N.L., Coble, P.A.: A Walsh–Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. J. Am. Stat. Assoc. **83**(404), 954–963 (1988)
Stoffer, D.S., Tyler, D.E., Wendt, D.A.: The spectral envelope and its applications. Stat. Sci. **15**(3), 224–253 (2000)
Weiß, C.H., Göb, R.: Measuring serial dependence in categorical time series. Adv. Stat. Anal. **92**(1), 71–89 (2008)
Weiß, C.H.: Empirical measures of signed serial dependence in categorical time series. J. Stat. Comput. Simul. **81**(4), 411–429 (2011)
Weiß, C.H.: Serial dependence of NDARMA process. Comput. Stat. Data Anal. **68**(1), 213–238 (2013)