Coherent Forecasting for Stationary Time Series of Discrete Data

Technical Report No. ASU/2014/8

Dated : 7^{th} August, 2014

Raju Maiti and Atanu Biswas

Applied Statistics Unit

Indian Statistical Institute

203, B. T. Road, Kolkata 700108, India.

E-mails: raju_r@isical.ac.in, atanu@isical.ac.in

Applied Statistics Unit

Indian Statistical Institute

Kolkata 700108

Coherent forecasting for stationary time series of discrete data

Raju Maiti and Atanu Biswas
Applied Statistics Unit, Indian Statistical Institute
203 B.T. Road, Kolkata - 700 108, India
E-mails: raju_r@isical.ac.in and atanu@isical.ac.in

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, mode forecasting can be used in order 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 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 illustrations. 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.

Key words: 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 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, however, 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 set up. But, in general, these models are not applicable in modeling categorical time series with finite number of categories.

Jacobs and Lewis (1978a, 1978b, 1978c), 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 set up – 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 for 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 like 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 details 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. In order 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 Section 2, different methods of coherent forecasting with some measures of forecasting accuracy are discussed in order to study the forecasting performance. Coherent forecasting for PAR(p), PMA(q) and PARMA(p,q) models are presented in Section 3. Coherent forecasting for MTD(p) and Logistic(p) models are discussed in Sections 4 and 5, respectively. Some extensive simulation results are presented in Section 6. In Section 7, a practical categorical data, namely infant sleep status data is 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 the 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. In order to use median predictor for categorical time series, 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 study 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

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

where $y_{(n+h)i}$ be the true i-th 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 the time series of real-valued data, the PRMSE and PMAE cannot be observed, especially 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, in order to study the forecasting accuracy for count and categorical data, we can always use another measure, namely percentage of true prediction (PTP) which can be defined as

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

In addition, we intend to propose some popular distance functions between true and predicted distributions as the measures of forecasting accuracy in order 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 $(k+1)$ categories $\{C_0, C_1, \ldots, C_k\}$, and 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 $\widehat{\mathbf{p}}_h$ denotes the h-step ahead forecasting distribution, then KSD, ED and MAD functions can be defined as

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

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

and

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

It is important to note that unlike KSD, the other two measures can be applied for 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. Therefore 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 confidence interval of Y_{n+h} given $\mathcal{Y}_n$. However, we can use some notion of confidence set in place of confidence 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*

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

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

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

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 confidence 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 $\{Y_t\}$ denotes 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 according as 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. However, under the assumption that the process is ordered categorical and $C_i = i$; $i = 0, 1, \ldots, k$, we derive the following moments and autocorrelation function for the PAR(p) process given in (3.2).

Let $\mathcal{Y}_t = (Y_t, Y_{t-1}, \ldots, Y_1)$, then the conditional expectation of Y_t given $\mathcal{Y}_{t-1}$ can be derived as

$$E(Y_t | \mathcal{Y}_{t-1}) = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + (1 - \phi_1 - \phi_2 - \cdots - \phi_p) \mu_\epsilon, \tag{3.5}$$

and the auto-covariance function (ACVF) of Y_t can be expressed as

$$\gamma_y(h) = Cov(Y_t, Y_{t-h}) = \phi_1 \gamma_y(|h-1|) + \phi_2 \gamma_y(|h-2|) + \cdots + \phi_p \gamma_y(|h-p|), \quad h \geq 1.$$

It follows that the auto-correlation function (ACF) is given by

$$\rho_y(h) = \phi_1 \rho_y(|h-1|) + \phi_2 \rho_y(|h-2|) + \cdots + \phi_p \rho_y(|h-p|), \text{ for } h \geq 1.$$

Combining the above equations for $h = 1, \ldots, p$, we have

$$\boldsymbol{\rho} = \boldsymbol{R}\boldsymbol{\phi},$$

where $\boldsymbol{\rho} = (\rho_y(1), \ldots, \rho_y(p))^T$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$, and $\boldsymbol{R} = ((\rho_{ij}))_{i,j=1}^p$ with $\rho_{ij} = \rho_y(|i-j|)$. Therefore we can obtain the YW estimators of $\boldsymbol{\phi}$ and is given by

$$\hat{\boldsymbol{\phi}}_{yw} = \widehat{\boldsymbol{R}}^{-1}\widehat{\boldsymbol{\rho}}.$$

In particular, ACF of a PAR(1) process can be derived as $\rho_y(h) = \phi^h$, and hence $\hat{\phi}_{yw} = \hat{\rho}_y(1)$.

It is important to notice that, if the time series is categorical, especially nominal categorical, where one cannot assign numerical values to the categories, the moments, autocorrelation function will not be defined. Although the auto-correlation 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). Those 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.

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*

$$\begin{aligned} p_h(i;\boldsymbol{\phi}) &= P(Y_{n+h} = C_i|\mathcal{Y}_n) \\ &= \eta_{h1}I(Y_n = C_i) + \cdots + \eta_{hp}I(Y_{n-p+1} = C_i) + (1 - \eta_{h1} - \cdots - \eta_{hp})p_i \\ &= \boldsymbol{\eta}_h^T\boldsymbol{e} + \left(1 - \boldsymbol{\eta}_h^T\mathbf{1}\right)p_i, \end{aligned} \tag{3.6}$$

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

$$\boldsymbol{\eta}_h = \boldsymbol{\Phi}^{h-1}\boldsymbol{\phi}, \tag{3.7}$$

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. □

In particular, if $C_i = i$, then we have the following corollaries:

Corollary 1. *The h-step ahead conditional mean can be written by*

$$\begin{aligned} E(Y_{n+h}|\mathcal{Y}_n) &= \eta_{h1}Y_n + \eta_{h2}Y_{n-1} + \cdots + \eta_{hp}Y_{n-p+1} + (1 - \eta_{h1} - \eta_{h2} - \cdots - \eta_{hp})\mu_\epsilon \\ &= \boldsymbol{\eta}_h^T \mathbf{Y}_{n:p} + \left(1 - \boldsymbol{\eta}_h^T \mathbf{1}\right) \mu_\epsilon, \end{aligned}$$

and the h-step ahead conditional variance can be expressed as

$$\begin{aligned} Var(Y_{n+h}|\mathcal{Y}_n) &= \eta_{h1}Y_n^2 + \eta_{h2}Y_{n-1}^2 + \cdots + \eta_{h1}Y_{n-p+1}^2 + (1 - \eta_{h1} - \cdots - \eta_{hp})\mathbf{m}^T\mathbf{p} \\ &\quad -\{\eta_{h1}Y_n + \cdots + \eta_{hp}Y_{n-p+1} + (1 - \eta_{h1} - \cdots - \eta_{hp})\mu_\epsilon\}^2 \\ &= \boldsymbol{\eta}_h^T \mathbf{Y}_{n:p}^2 + \left(1 - \boldsymbol{\eta}_h^T \mathbf{1}\right) \mathbf{m}^T\mathbf{p} - \left\{\boldsymbol{\eta}_h^T \mathbf{Y}_{n:p} + \left(1 - \boldsymbol{\eta}_h^T \mathbf{1}\right) \mu_\epsilon\right\}^2, \end{aligned}$$

where $\mathbf{1} = (1, 1, \ldots, 1)^T$ *and* $\mathbf{Y}_{n:p} = (Y_n, Y_{n-1}, \ldots, Y_{n-p+1})^T$ *denote p-dimensional vectors, and* $\mathbf{m} = (0^2, 1^2, \ldots, k^2)^T$ *and* $\mathbf{p} = (p_0, p_1, \ldots, p_k)^T$ *are all* $(k+1)$*-dimensional vectors.*

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

Proposition 1. *Under the above set up, 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 the 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). Using the above proposition, it can be obtained that

$$\lim_{h\to\infty} E(Y_{n+h}|\mathcal{Y}_n) = \mu_y \text{ and } \lim_{h\to\infty} Var(Y_{n+h}|\mathcal{Y}_n) = \sigma_y^2,$$

where σ_y^2 is the marginal variance of Y_t.

3.2.1 *Confidence interval for Pegram's AR(p) model when parameters are estimated*

If the parameters are known, it is straightforward to obtain $p_h(i;\boldsymbol{\phi})$. However, in practice parameters are to be estimated using different methods of estimation. Suppose $\widehat{\boldsymbol{\phi}}_{yw}$ is the YW estimator of $\boldsymbol{\phi}$ under the assumption that $C_i = i$, then it can be shown that $\sqrt{n}(\widehat{\boldsymbol{\phi}}_{yw} - \boldsymbol{\phi}) \overset{a}{\sim} N(0, V)$ for some covariance matrix V. Proof is exactly same as in case of Box and Jenkins' AR(p) process (see Brockwell and Davis, 2002). Now, using delta-method, the estimated h-step ahead probability mass function, $p_h(i;\widehat{\boldsymbol{\phi}}_{yw})$ is asymptotically normal with mean $p_h(i;\boldsymbol{\phi})$

and variance $\sigma_h^2(i,\boldsymbol{\phi})$, where $\sigma_h^2(i,\boldsymbol{\phi}) = \bigtriangledown g(\boldsymbol{\phi})^T.(\frac{V}{n}).\bigtriangledown g(\boldsymbol{\phi})$ (see Freeland and McCabe (2004)). In particular for the PAR(1) model,

$$\sigma_h^2(i,\phi) = n^{-1}\{g'(\phi)\}^2(1-\phi^2) = n^{-1}\left\{h\phi^{h-1}\left(I(Y_n=i)-p_i\right)\right\}^2(1-\phi^2).$$

Therefore, $100(1-\alpha)\%$ asymptotic confidence interval for $p_h(i,\boldsymbol{\phi})$ can be obtained as $p_h(i,\widehat{\boldsymbol{\phi}}_{yw}) \mp \tau_{\alpha/2}\sigma_h(i,\widehat{\boldsymbol{\phi}}_{yw})$, where $\tau_{\alpha/2}$ is the upper $\alpha/2$-th point of standard normal distribution.

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 $t \in 0,\pm1,\pm2,\ldots$, the conditional probability function takes the form

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

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, 1978b). Under the assumption that $C_i = i$, its ACVF at lag h is given by

$$\gamma(h) = \begin{cases} \sum_{r=0}^{q-h}\theta_r\theta_{r+h}, & \text{if } 0 \leq h \leq q \\ 0, & \text{if } h > q. \end{cases} \tag{3.8}$$

Based on this ACVF, one can obtain the YW estimates of the model parameters.

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 $1 \leq h \leq q$ and $l = q - 1$,

$$\begin{aligned} p_h(i) &= P(Y_{n+h} = C_i | Y_n = C_{i_0}, \ldots, Y_{n-l} = C_{i_l}) \\ &= \frac{\sum_{r_h=0}^{q} \sum_{r_0=0}^{q} \cdots \sum_{r_l=0}^{q} \theta_{r_h} \theta_{r_0} \cdots \theta_{r_l} P(\epsilon_{n+h-r_h} = C_i, \epsilon_{n-r_0} = C_{i_0}, \ldots, \epsilon_{n-l-r_l} = C_{i_l})}{\sum_{r_0=0}^{q} \cdots \sum_{r_l=0}^{q} \theta_{r_0} \cdots \theta_{r_l} P(\epsilon_{n-r_0} = C_{i_0}, \ldots, \epsilon_{n-l-r_l} = C_{i_l})}, \end{aligned} \tag{3.9}$$

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 $h \geq 2$. In order 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 $0 < h \leq q$, we have

$$\begin{aligned} P(Y_{n+h} = C_i | Y_n = C_j) &= \frac{P(Y_{n+h} = C_i, Y_n = C_j)}{P(Y_n = C_j)} \\ &= \left(\sum_{r=0}^{q-h} \theta_r \theta_{r+h} \right) \{ I(i = j) - p_i \} + p_i, \end{aligned} \tag{3.10}$$

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

To study the difference in using the conditional distribution of Y_{n+h} given Y_n presented in (3.10) over the use of true forecasting distribution given in (3.9), we carry out one simulation study for the PMA(2) process for different possible choices of the model parameters. We reported the results for $n = 500$ with model parameters $(\theta_0, \theta_1, \theta_2) = (0.2, 0.6, 0.2)$ and the marginal distribution $\boldsymbol{p} = (0.2, 0.1, 0.5, 0.15, 0.05)$ defined on the state space $S = \{0, 1, 2, 3, 4\}$. Based on the simulated data we obtain the exact forecasting distribution using the formula given in Appendix C and the conditional distribution of Y_{n+h} given the present observation Y_n given in (3.10). The fitted forecasting distribution and the fitted conditional distribution are presented in Figure 1. As one can see no significant difference is visualized. Therefore one can use the conditional distribution given in (3.10) as an alternative to the actual forecasting

distribution given in (3.9) which would be very difficult 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

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

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. Under the assumption that $C_i = i$, ACF of the PARMA(1,1) model is given by

$$\rho(h) = \begin{cases} \phi_1 + \theta_0 \theta_1 & \text{if} \quad h = \pm 1, \\ \phi_1^h & \text{if} \quad h = \pm 2, \pm 3, \ldots. \end{cases}$$

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

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

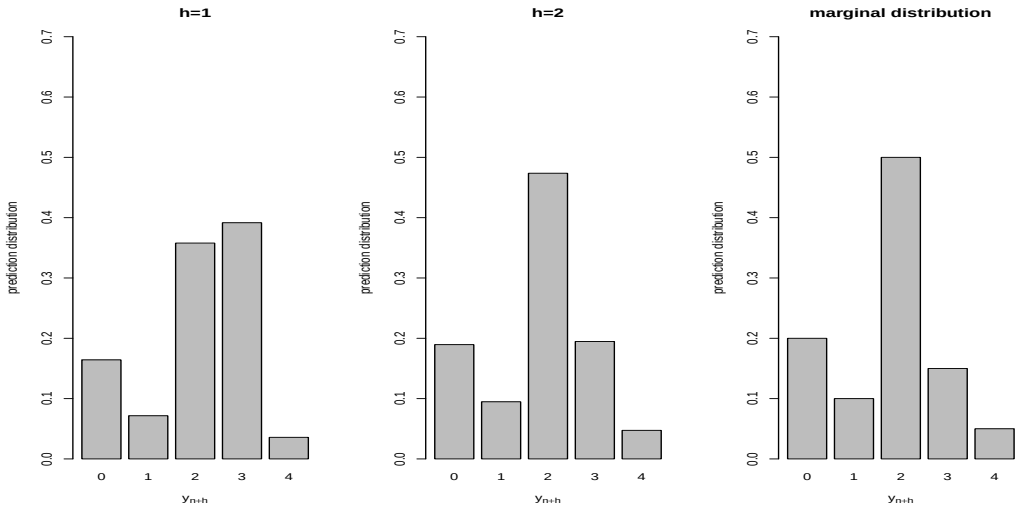
and for $h > 1$,

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

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

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

Forecasting distribution of Y_{n+h} given $Y_n, Y_{n-1}, \cdots$



Conditional distribution of Y_{n+h} given Y_n

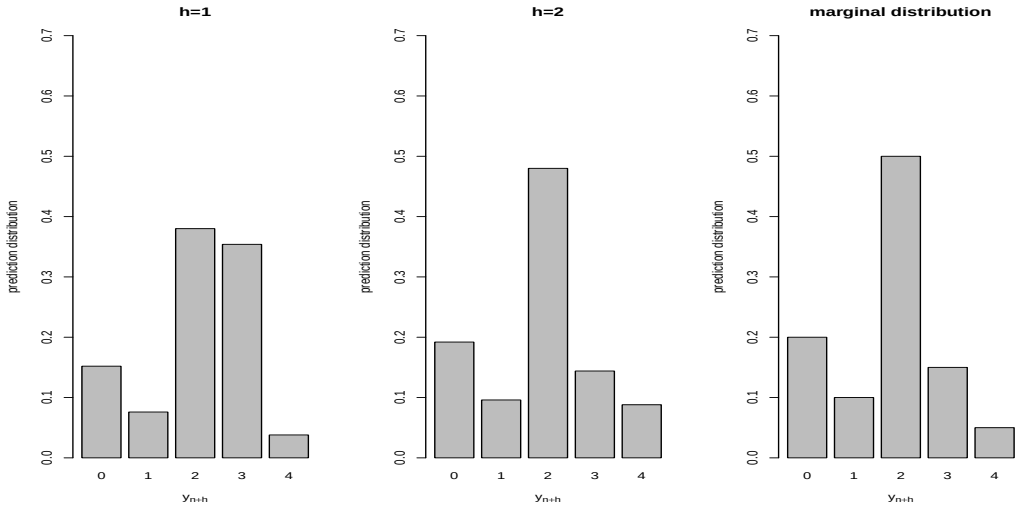


Figure 1: h-step ahead forecasting and conditional distributions for the PMA(2) process with $(\theta_0, \theta_1, \theta_2) = (0.2, 0.6, 0.2)$ and marginal distribution $\mathbf{p} = (0.2, 0.1, 0.5, 0.15, 0.05)$.

where $\mathbf{e} = (I(i_0 = i), I(i_1 = i), \ldots, I(i_{p-1} = i))^T$ 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 (3.7). Similarly, for the PARMA(p,2) model and for $h = 1$ we have,

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

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). The MTD model 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 is assumed 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

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

Similarly, three-step ahead forecasting distribution is given by

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

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 in order to build regression models for categorical time series. Here we provide a brief description of the multinomial logistic regression 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 $0, 1, 2 \ldots k$ where the "first" category is assigned the integer value of 0, the "second" category is assigned the integer value of 1 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 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

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

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*

In order 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,

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

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 in order to obtain 2-step ahead forecasting i.e. $Y_{(n+2)}$ and repeat this process for $h = 2, 3, \ldots,$ to obtain the h-step ahead forecasting value. 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 Section 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

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

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

In order to study the finite sample behaviors of the proposed forecasting measures, namely PTP, KSD, ED and MAD, and the interval forecasting measure defined in Section 2, and to facilitate model comparison through the Akaike information criterion (AIC) and 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

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

(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 **M1**, **M2** and **M3**. Five sample sizes are explored, samples of sizes 100 and 300 are used to study the small sample properties; samples of sizes 500 and 1000 are used to get an idea about the moderate sample properties, and samples of size 5000 are used to study the large sample properties. For a fixed sample size n, we repeated the process 1000 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**, whereas Tables 2 and 3 summarize the results based on samples from **M2** and **M3**, respectively. 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 5000 replications are reported in Table 4, 5, and 6. As we can see from the tables, for all the

Table 1: *Percentage of times AIC and BIC select the correct model where data is generated from* **M1**.

Sample size (n)	AIC			BIC		
	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
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
1000	100	0	0	100.0	0	0
5000	100	0	0	100.0	0	0

cases, the measures KSD, ED and MAD are increasing in h. 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 Tables 4, 5, 6). 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.

Here we repeated the previous exercise, where we simulated samples of size 150 from all the three cases **M1**, **M2**, and **M3** in order 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 are presented in Tables 7, 8 and 9. 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, 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 minutes. In this Section, we consider

Table 2: *Percentage of times AIC and BIC select the correct model where data is generated from* **M2**.

Sample size (n)	AIC			BIC		
	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
100	0	100	0	10	90	0
300	0	100	0	0	100	0
500	0	100	0	0	100	0
1000	0	100	0	0	100	0
5000	0	100	0	0	100	0

Table 3: *Percentage of times AIC and BIC select the correct model where data is generated from* **M3**.

Sample size (n)	AIC			BIC		
	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
100	0	0	100	0	0	100
300	0	0	100	0	0	100
500	0	0	100	0	0	100
1000	0	0	100	0	0	100
5000	0	0	100	0	0	100

Table 4: *Values of forecasting measures PTP, KSD, ED and MAD for varying h where the data generating model is* **M1**.

h-step	PTP(h)			KSD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)		
	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
1	83.68	83.36	83.62	0.0185	0.0257	0.0247
2	72.26	71.44	70.76	0.0302	0.0408	0.0468
3	62.55	61.22	60.78	0.0376	0.0487	0.0507
4	53.50	52.80	53.00	0.0425	0.0522	0.0512
5	48.68	46.23	45.83	0.0456	0.0524	0.0527
6	42.52	41.14	39.77	0.0489	0.0516	0.0546
	ED($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)			MAD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)		
1	0.0294	0.0419	0.0418	0.0221	0.0329	0.0349
2	0.0491	0.0671	0.0721	0.0370	0.0529	0.0609
3	0.0622	0.0812	0.0791	0.0473	0.0638	0.0628
4	0.0708	0.0880	0.0879	0.0542	0.0685	0.0725
5	0.0765	0.0903	0.0932	0.0586	0.0693	0.0793
6	0.0803	0.0902	0.0901	0.0616	0.0686	0.0826

Table 5: *Values of forecasting measures PTP, KSD, ED and MAD for varying h where the data generating model is* **M2**.

h-step	PTP(h)			KSD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)		
	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
1	84.79	85.90	84.86	0.0671	0.0143	0.0156
2	73.73	74.84	73.93	0.1039	0.0245	0.0246
3	65.65	67.90	65.67	0.1296	0.0325	0.0339
4	59.39	60.53	59.27	0.1455	0.0391	0.0427
5	54.50	55.71	54.59	0.1548	0.0446	0.0502
6	50.64	51.62	51.54	0.1594	0.0491	0.0566
	ED($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)			MAD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)		
1	0.0780	0.0198	0.0223	0.0586	0.0154	0.0171
2	0.1268	0.0327	0.0356	0.0961	0.0255	0.0278
3	0.1596	0.0426	0.0495	0.1212	0.0334	0.0383
4	0.1817	0.0506	0.0627	0.1381	0.0398	0.0479
5	0.1959	0.0572	0.0730	0.1493	0.0449	0.0559
6	0.2042	0.0627	0.0814	0.1559	0.0494	0.0624

Table 6: *Values of forecasting measures PTP, KSD, ED and MAD for varying h where the data generating model is* **M3**.

h-step	PTP(h)			KSD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)		
	PAR(1)	MTD(1)	Logistic(1)	PAR(1)	MTD(1)	Logistic(1)
1	92.49	92.25	93.35	0.0631	0.0152	0.0130
2	86.37	86.59	87.46	0.1003	0.0238	0.0219
3	80.94	81.17	82.89	0.1219	0.0360	0.0289
4	76.14	76.48	77.09	0.1351	0.0361	0.0345
5	72.04	72.08	72.67	0.1429	0.0406	0.0392
6	68.11	67.98	68.81	0.1470	0.0443	0.0430
	ED($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)			MAD($\mathbf{p}_h$, $\widehat{\mathbf{p}}_h$)		
1	0.0737	0.0205	0.0166	0.0586	0.0159	0.0131
2	0.1176	0.0288	0.0262	0.0923	0.0227	0.0213
3	0.1450	0.0358	0.0338	0.1111	0.0289	0.0277
4	0.1626	0.0419	0.0401	0.1226	0.0342	0.0331
5	0.1738	0.0472	0.0455	0.1302	0.0386	0.0378
6	0.1804	0.0516	0.0501	0.1350	0.0423	0.0416

Table 7: $100(1-\alpha)\%$ HPP set of Y_{n+h} given Y_n for varying h with cardinality of the set, where data is generated from **M1** where $\alpha = 0.2$.

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

Table 8: $100(1-\alpha)\%$ HPP set of Y_{n+h} given Y_n for varying h with cardinality of the set, where data is generated from **M2** where $\alpha = 0.2$.

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

Table 9: $100(1-\alpha)\%$ HPP set of Y_{n+h} given Y_n for varying h with cardinality of the set where data is generated from **M3** where $\alpha = 0.2$.

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

one such single time series from the first group. The raw data is presented in Figure 2.

During minute t, the infant's sleep status was recorded in 6 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, $\mathbf{Y}_t$ has 5 components, i.e. $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, Y_{t3}, Y_{t4}, Y_{t5})$, if we want to fit the logistic regression model, . 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 equation (5.2). Note that partial likelihood estimates of 30 parameters based on the data of size 128 may not be so reliable. Therefore, in order 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. Since the states have an order, therefore after combining the quite states and active states, one obvious scaling made in Stoffer et al. (2000) is

$$\text{qt} \equiv 0, \quad \text{qh} \equiv 0, \quad \text{tr} \equiv 1, \quad \text{al} \equiv 2, \quad \text{ah} \equiv 2, \quad \text{aw} \equiv 3. \tag{7.1}$$

The proportion of times spent by an infant in the combined sleep status $0, 1, 2,$ and 3 are 0.414, 0.008, 0.539, and 0.039, respectively. This indicates that infant spent maximum time in active sleep. Based on the scaling (7.1), we observed the values of autocorrelation function (ACF) and partial autocorrelation function (PACF) for various lag values, and plotted it in Figure 2. Note that Figure 2 consists of three plots: the top one is the time series plot of the scaled data and middle one displays the sample ACF, which is decreasing in its lag values. The bottom one displays the sample PACF, in which only the first lag appears to be significant. Hence there is a strong evidence in favor of AR(1) model.

However, it is important to notice that although the infant sleep status data is 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 $C_1 < C_2 < C_3 < C_4$, and cannot specify the values of C_1, C_2, C_3, C_4. Hence some alternative measures of serial association, which do not depend on the numerical scaling of the categories, should be used in order 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 $\widehat{A}_{\nu}^{(\tau)}(h)$ 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 details 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 10. 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 approximately equal. 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 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 $n = 200, 1000, 10000$ 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 Figure 4. As we can see from Figure 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 as $(0.414, 0.008, 0.539, 0.0391)$. Similarly the transition probability matrix

Table 10: Estimated values of $\kappa(h)$, $v(h)$, $A_{\nu}^{(\tau)}(h)$ and Cohen's κ based partial auto-correlation ($\rho_{\mathrm{p}}(h)$) for the infant sleep status data.

Lag h	$\widehat{\kappa}(h)$	$\widehat{v}(h)$	$\sqrt{\widehat{A}_{\nu}^{(\tau)}(h)}$	$\widehat{\rho}_{\mathrm{p}}(h)$
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

(tpm) $\mathbf{Q}$ associated with MTD(1) model is estimated as

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

In order to fit the Logistic regression model, we used the set up discussed in equation (5.1) in Section 5. Note that, after combining the states the data has 4 categories and hence $\mathbf{Y}_t$ has 3 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 Figure 5. As one can see, there is a decreasing pattern in the first and last plots in Figure 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

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

After fitting the above models, we obtained the AIC and BIC for all the three models and presented it in Table 11. 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 11. 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.

Table 11: 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

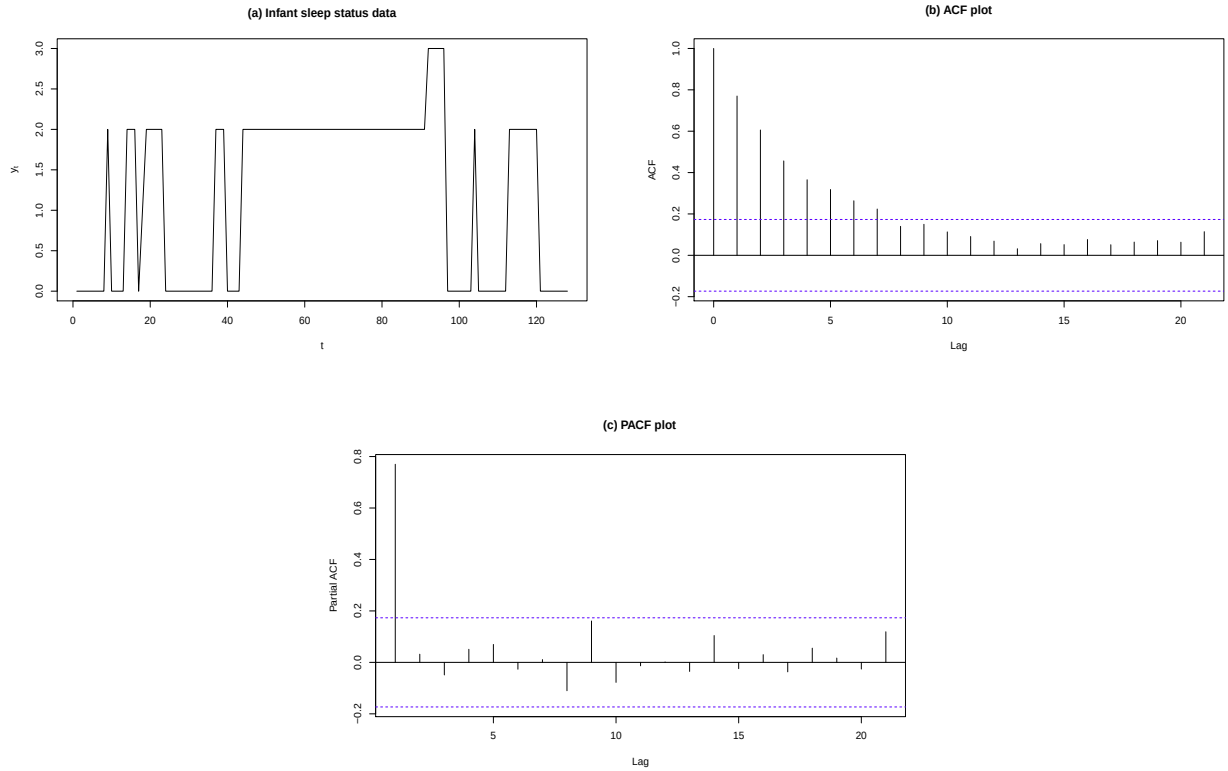


Figure 2: Plot of ACF and PACF for the infant sleep status data.

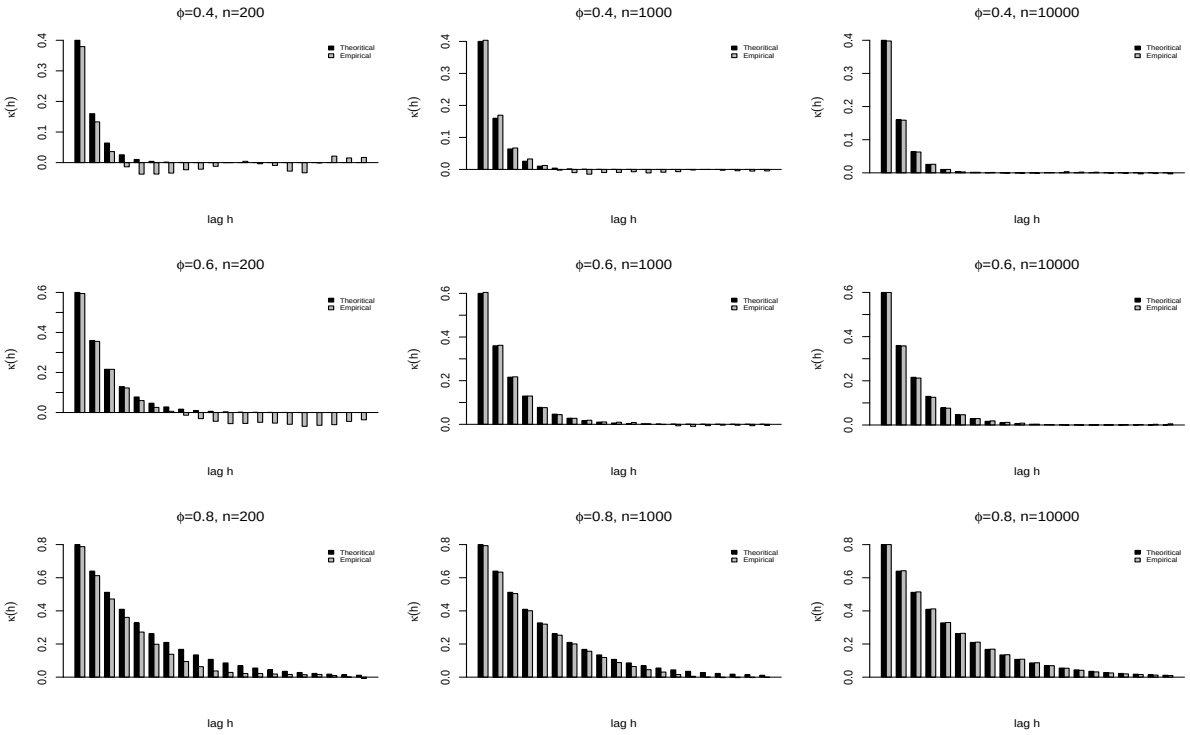


Figure 3: Theoretical and empirical values of Cohen's κ for various lag values h. Here samples are generated from PAR(1) with number of categories 4, for mixing parameter $\phi = 0.4, 0.6, 0.8$ and sample sizes $n = 200, 1000, 10000$ with the common marginal distribution $\mathbf{p} = (0.414, 0.008, 0.539, 0.039)$.

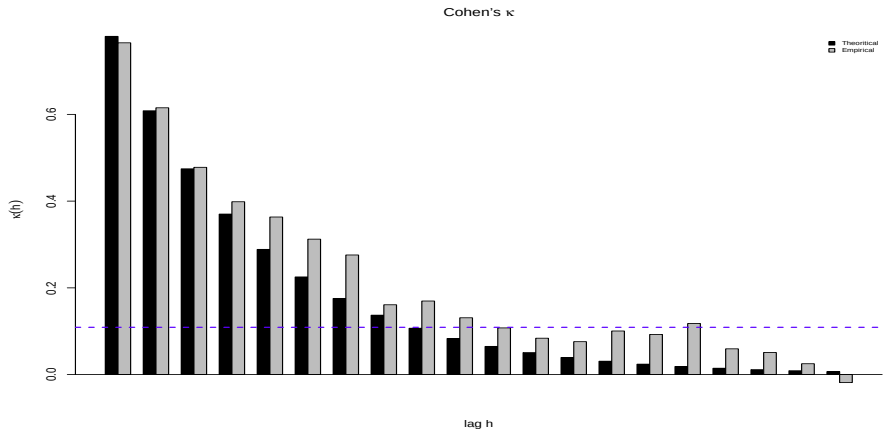


Figure 4: Plot of Cohen's κ for varying lag values for the infant sleep data.

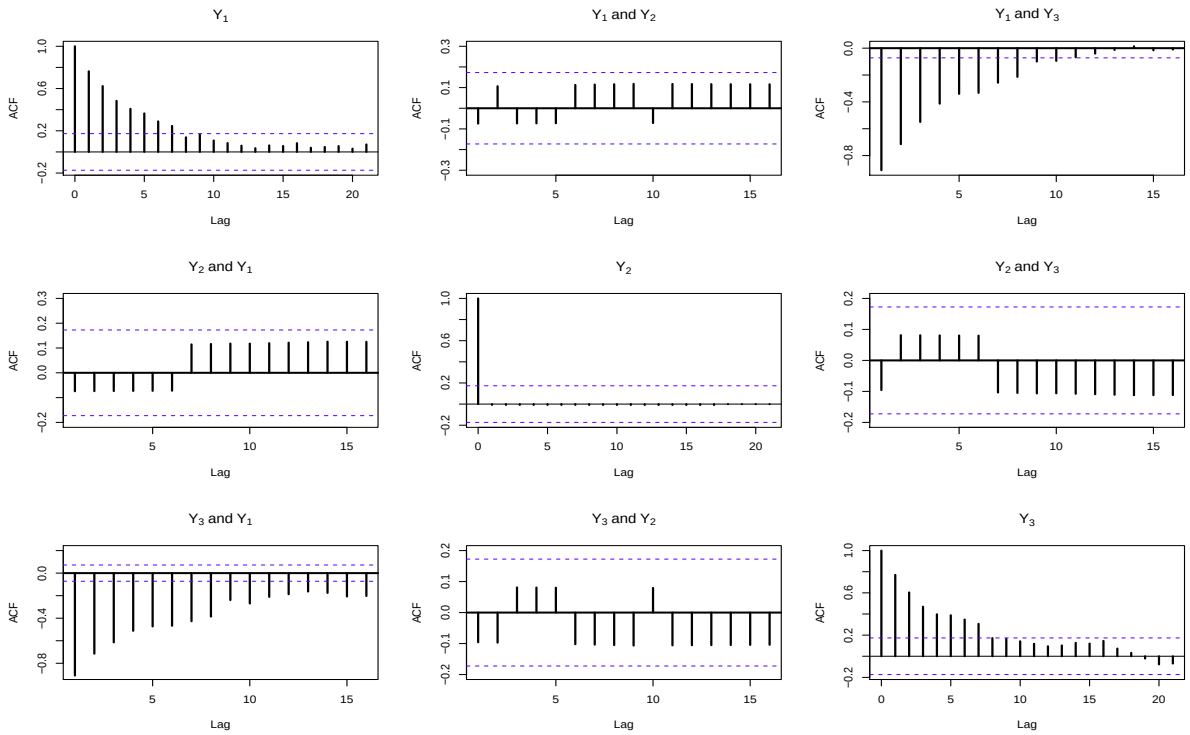


Figure 5: Sample autocorrelation and cross-correlation functions for the infant sleep status data.

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 Section 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, PMAE cannot be used. Therefore in order 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 count data and categorical data both (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 are 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 is 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.

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

References

Agresti, A. (2002). *Categorical data analysis*. John Wiley & Sons, New Jersey. 17

Al-Osh, M. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8(3):261–275. 3

Alzaid, A. and Al-Osh, M. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, 27:314–324. 3

Berchtold, A. and Raftery, A. E. (2002). The mixture transition distribution model for high-order Markov chains and non-gaussian time series. *Statistical Science*, 17(3):328–356. 3

Biswas, A. and Guha, A. (2009). Time series analysis of categorical data using auto-mutual information. *Journal of Statistical Planning and Inference*, 139(9):3076–3087. 7, 29

Biswas, A. and Song, P. X.-K. (2009). Discrete-valued ARMA processes. *Statistics & Probability Letters*, 79(17):1884–1889. 3, 7, 8, 11, 13, 29

Brockwell, P. J. and Davis, R. A. (2002). *Time series: theory and methods*. Springer. 10

Bu, R. and McCabe, B. (2008). Model selection, estimation and forecasting in INAR(p) models: A likelihood-based Markov chain approach. *International journal of forecasting*, 24(1):151–162. 3

Carruth, J., Tygert, M., and Ward, R. (2012). A comparison of the discrete kolmogorov-smirnov statistic and the euclidean distance. *arXiv preprint arXiv:1206.6367*. 6

Fokianos, K. and Kedem, B. (2003). Regression theory for categorical time series. *Statistical science*, 18(3):357–376. 3, 17, 18, 30

Freeland, R. K. and McCabe, B. P. (2004). Forecasting discrete valued low count time series. *International Journal of Forecasting*, 20(3):427–434. 2, 4, 11

Jacobs, P. A. and Lewis, P. A. (1978a). Discrete time series generated by mixtures. I: Correlational and runs properties. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):94–105. 3, 11

Jacobs, P. A. and Lewis, P. A. (1978b). Discrete time series generated by mixtures II: asymptotic properties. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):222–228. 3, 11

Jacobs, P. A. and Lewis, P. A. (1978c). Discrete time series generated by mixtures III: Autoregressive processes (DAR(p)). Technical report, Monterey, California. Naval Postgraduate School. 3, 10

Jacobs, P. A. and Lewis, P. A. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36. 3, 7, 10, 13

Jung, R. C. and Tremayne, A. R. (2006). Coherent forecasting in integer time series models. *International Journal of Forecasting*, 22(2):223–238. 3

McKenzie, E. (1985). Some simple models for discrete variate time series. *JAWRA Journal of the American Water Resources Association*, 21(4):645–650. 3

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, 20(4):822–835. 3

Pegram, G. (1980). An autoregressive model for multilag Markov chains. *Journal of Applied Probability*, 17(2):350–362. 3, 7, 10

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3):528–539. 3, 7, 15, 30

Rao, A. and Bhimsankaram, P. (2000). *Linear Algebra*. Hindustan, India. 33

Silva, N., Pereira, I., and Silva, M. E. (2009). Forecasting in INAR(1) model. *REVSTAT – Statistical Journal*, 7(1):119–134. 3

Stoffer, D. S., Scher, M. S., Richardson, G. A., Day, N. L., and Coble, P. A. (1988). A walshfourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *Journal of the American Statistical Association*, 83(404):954–963. 20

Stoffer, D. S., Tyler, D. E., and Wendt, D. A. (2000). The spectral envelope and its applications. *Statistical Science*, 15(3):224–253. 24

Weiß, C. H. (2011). Empirical measures of signed serial dependence in categorical time series. *Journal of Statistical Computation & Simulation*, 81(4):411–429. 25

Weiß, C. H. (2013). Serial dependence of NDARMA process. *Computational Statistics & Data Analysis*, 68(1):213–238. 25

Weiß, C. H. and Göb, R. (2008). Measuring serial dependence in categorical time series. *AStA Advances in Statistical Analysis*, 92(1):71–89. 3, 7, 8, 9, 25

9 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 2-step ahead conditional distribution is given by

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

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 (3.6).

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 Rao and Bhimsankaram (2000)).

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

$$A^k = VJ^kV^{-1}$$

and, since J is block diagonal,

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

Now a standard result on the k^{th} 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,

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

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 $h = 2$,

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

where

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

and

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

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