



A p th-order random coefficients mixed binomial autoregressive process with explanatory variables

Han Li^{1,2} · Zijian Liu² · Kai Yang² · Xiaogang Dong² · Wenshan Wang²

Received: 6 January 2023 / Accepted: 24 July 2023 / Published online: 8 August 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

To capture the higher-order autocorrelation structure for finite-range integer-valued time series of counts, and to consider the driving effect of covariates on the underlying process, this paper introduces a p th-order random coefficients mixed binomial autoregressive process with explanatory variables. The basic probabilistic and statistical properties of the model are discussed. Conditional least squares and conditional maximum likelihood estimators, as well as their asymptotic properties of the estimators are obtained. Moreover, the existence test of explanatory variables are well addressed using a Wald-type test. Forecasting problem is also considered. Finally, some numerical results of the estimators and a real data example are presented to show the performance of the proposed model.

Keywords Finite-range integer-valued time series · Binomial autoregressive model · Random coefficient model · Explanatory variables · Forecasting

1 Introduction

As an important branch of time series analysis, integer-valued time series has attracted more and more attention in recent years. This kind of data is widely used in various fields of our daily life. For example, the annual counts of world major earthquakes (Wang et al. 2014; Yang et al. 2018), the monthly number of cases of an infectious disease (Pedeli et al. 2015; Yang et al. 2022), and the number of areas in which an infectious disease occurs per week (Ristić et al. 2016; Chen et al. 2019), among others. According to the different value ranges of the observed data, such data can be divided into two categories. The first category is integer-valued time series data that take values on the set of natural numbers $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. The

✉ Kai Yang
yangkai@ccut.edu.cn

¹ School of Science, Changchun University, Changchun 130012, China

² School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China

well-known integer-valued autoregressive (INAR) model (Al-Osh and Alzaid 1987) is a typical representative on modelling such data. The second category is integer-valued time series with a finite-range support, say $\mathbb{S} = \{0, 1, \dots, N\}$. To model finite-range integer-valued time series of counts (McKenzie 1985) proposed the first-order integer-valued binomial autoregressive (BAR(1)) model, which is defined as follows:

$$X_t = \alpha \circ X_{t-1} + \beta \circ (N - X_{t-1}),$$

where $\alpha, \beta \in (0, 1)$, “ \circ ” is the binomial thinning operator proposed by Steutel and van Harn (1979). Let X be an integer-valued random variable and $\alpha \in (0, 1)$, the binomial thinning is defined as

$$\alpha \circ X := \begin{cases} \sum_{i=1}^X B_i, & \text{if } X > 0, \\ 0, & \text{if } X = 0, \end{cases} \quad (1.1)$$

where $\{B_i\}$ is a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables satisfying $P(B_i = 1) = 1 - P(B_i = 0) = \alpha$, which is also independent of X . All thinnings are performed independently for the BAR(1) model.

Since the seminal work by McKenzie (1985), modelling and inference for finite-range time series of counts have received considerable attention. Brännäs and Nordström (2006) generalized the BAR(1) model by replacing N with N_t to present an econometric model to account for the tourism accommodation impact of arranging festivals or special events in many cities. Weiß (2009b) generalized the BAR(1) model to p th-order and develop a BAR(p) model. Weiß and Pollett (2014) proposed a binomial autoregressive process with density dependent thinning. Möller et al. (2016) developed a self-exciting threshold binomial autoregressive (SET-BAR(1)) process. Yang et al. (2018) contributed the empirical likelihood inference for the SET-BAR(1) model, and addressed the problem of estimating the threshold parameter of the SET-BAR(1) model. Zhang et al. (2020) proposed a multinomial autoregressive model for finite-range integer-valued time series with more than two states. Nik and Weiß (2021) developed a binomial smooth-transition autoregressive models for time series of bounded counts. For recent achievements and applications of binomial autoregressive models, we refer the readers to Weiß (2009a), Scotto et al. (2014), Chen et al. (2020), Kang et al. (2021), Zhang et al. (2022) and among others.

Researchers found that regression models for time series of counts are becoming increasingly often applied (Brännäs 1995). However, the binomial autoregressive models mentioned above ignore the effect of exogenous variables on the observed data. To address this problem in the area of infinite-range time series of counts, scholars have made different attempts. To make the analyzed models more applicable, Freeland and McCabe (2004a) introduced explanatory variables into the parameters of first-order INAR model via two different kinds of link functions. Enciso-Mora et al. (2009) proposed an INAR(p) process with explanatory variables both in the autoregressive coefficients and the expectation of innovation. Ding and Wang (2016), and Wang (2020) successively studied the empirical likelihood inferences and variable selection problems for first-order Poisson integer-valued autoregressive

model with covariables. Yang et al. (2021) confirmed the existence of a nonlinear relationship of climate covariates on crime cases, and further suggested a random coefficients integer-valued threshold autoregressive processes driven by logistic regression. This research further expands the study of the INAR model with covariables and enhances the applicability of the model.

To capture the impact of covariates on the finite-range time series of counts, Wang et al. (2021) developed a first-order covariates-driven binomial AR (CDBAR(1)) process. Zhang and Wang (2023) developed a binomial AR(1) process with autoregressive coefficient driven by a bivariate dependent autoregressive process with covariables. However, the two models are both first-order models. To the best of our knowledge, there is no literature discussing the high-order modelling for finite-range integer-valued time series of counts with explanatory variables. In fact, a high-order model for time series of counts is indeed very important, which is recognized by many scholars (see, e.g., Zhu and Joe (2006), Weiß (2009b), Yang et al. (2023) and among others). In this study, we aim to make a contribution towards this direction.

The remainder of the paper is organized as follows. In Sect. 2, we introduce the definition and basic properties of the p th-order random coefficients mixed binomial autoregressive process with explanatory variables, and denote the proposed model as RCMBAR(p)-X process. In Sect. 3, we discuss the parameter estimation problem via two different methods, the asymptotic properties of the estimators are also provided. In Sect. 4, we develop a Wald-type test to address the testing problem for the existence of explanatory variables. In Sect. 5, forecasting problem for the proposed model is addressed. In Sect. 6, we conduct some simulation studies to show the performances of the proposed methods. In Sect. 7, we apply the proposed method to the weekly rainfall data set in Germany. Some concluding remarks are given in Sect. 8. All proofs are postponed to the Appendixes.

2 Definition and basic properties of the RCMBAR(p)-X process

In this section, we first introduce the definition of the p th-order random coefficients mixed binomial autoregressive process with explanatory variables, and then give some important properties of it. The definition of RCMBAR(p)-X process is given as follows:

Definition 1 A sequence of integer-valued random observations $\{X_t\}_{t \in \mathbb{Z}}$ is said to follow a p th-order random coefficients mixed integer-valued binomial autoregressive process with explanatory variables, if X_t satisfies the recursion

$$X_t = \begin{cases} \alpha_t \circ X_{t-1} + \beta_t \circ (N - X_{t-1}), & w.p. \phi_1, \\ \alpha_t \circ X_{t-2} + \beta_t \circ (N - X_{t-2}), & w.p. \phi_2, \\ \dots \\ \alpha_t \circ X_{t-p} + \beta_t \circ (N - X_{t-p}), & w.p. \phi_p, \end{cases} \quad (2.2)$$

where “o” is the binomial thinning operator defined in (1.1), $N \in \mathbb{N}$ is a predetermined upper limit of the range, the weights $\phi_1, \phi_2, \dots, \phi_p \in (0, 1)$, $\sum_{i=1}^p \phi_i = 1$, “w.p.” stands for with probability, $\alpha_t, \beta_t \in (0, 1)$ are the autoregressive coefficients satisfying

$$\log\left(\frac{\alpha_t}{1 - \alpha_t}\right) = \mathbf{Z}_t^\top \boldsymbol{\delta}_1, \quad \log\left(\frac{\beta_t}{1 - \beta_t}\right) = \mathbf{Z}_t^\top \boldsymbol{\delta}_2, \tag{2.3}$$

where $\boldsymbol{\delta}_i := (\delta_{i,0}, \delta_{i,1}, \dots, \delta_{i,q})^\top$, $i = 1, 2$, are the regression coefficients, $\{\mathbf{Z}_t := (1, Z_{t,1}, \dots, Z_{t,q})^\top\}$ is a sequence of explanatory variables with constant mean vector and covariance matrix. For a fixed \mathbf{Z}_t , the thinning operations at time t are performed independently of each other.

As is seen in Definition 1, the RCMBAR(p)-X process is actually a mixture integer-valued autoregressive model with fixed weights. X_t equals $\alpha_t \circ X_{t-i} + \beta_t \circ (N - X_{t-i})$ with probability ϕ_i , $i = 1, 2, \dots, p$. Furthermore, the autoregressive coefficients α_t and β_t shared the randomness and flexibility via a logistic structure with covariates. Obviously, Definition 1 includes the covariates-driven binomial AR(1) process of Wang et al. (2021) as a special case when $p = 1$. The RCMBAR(p)-X model reduces to the binomial AR(p) model of Weiß (2009b) when $\delta_{ij} = 0$ for $i = 1, 2$ and $j = 1, 2, \dots, q$.

Denote by $\{\mathbf{D}_t\}$ a sequence of i.i.d. multinomial random variables with parameters ϕ_1, \dots, ϕ_p , i.e., $\mathbf{D}_t = (D_{t,1}, D_{t,2}, \dots, D_{t,p})^\top \sim \text{MULT}(1; \phi_1, \dots, \phi_p)$, then model (2.2) can be equivalently rewritten in the following form:

$$X_t = \sum_{i=1}^p D_{t,i} (\alpha_t \circ X_{t-i} + \beta_t \circ (N - X_{t-i})), \tag{2.4}$$

where \mathbf{D}_t is independent of all X_s , $\alpha_t \circ X_{t-s}$ and $\beta_t \circ (N - X_{t-s})$ with $s < t$,

$$\alpha_t = \frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}, \quad \beta_t = \frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2)}{1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2)}, \tag{2.5}$$

are implied by (2.3). It follows by the expression (2.4) that the conditional probability of X_t conditional on X_{t-i} ($i = 1, 2, \dots, p$) and \mathbf{Z}_t fixed is given by

$$\begin{aligned} &P(X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}, \mathbf{Z}_t) \\ &= \sum_{i=1}^p \phi_i \sum_{m=a}^b \binom{x_{t-i}}{m} \binom{N - x_{t-i}}{x_t - m} \alpha_t^m (1 - \alpha_t)^{x_{t-i} - m} \beta_t^{x_t - m} (1 - \beta_t)^{N - x_{t-i} - x_t + m} \\ &= \sum_{i=1}^p \phi_i \sum_{m=a}^b \binom{x_{t-i}}{m} \binom{N - x_{t-i}}{x_t - m} \frac{\exp(m \mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^{x_{t-i}}} \frac{\exp((x_t - m) \mathbf{Z}_t^\top \boldsymbol{\delta}_2)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2))^{N - x_{t-i}}}, \end{aligned} \tag{2.6}$$

where $a = \max\{0, x_t + x_{t-i} - N\}$, $b = \min\{x_t, x_{t-i}\}$. The above conditional probability can be used to derive the conditional likelihood for the RCMBAR(p)-X

process. Furthermore, the conditional expectation and conditional variance are given by

$$E(X_t | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) = \sum_{i=1}^p \phi_i \left(\frac{\exp(\mathbf{Z}_t^\top \delta_1)}{1 + \exp(\mathbf{Z}_t^\top \delta_1)} X_{t-i} + \frac{\exp(\mathbf{Z}_t^\top \delta_2)}{1 + \exp(\mathbf{Z}_t^\top \delta_2)} (N - X_{t-i}) \right),$$

and

$$\begin{aligned} \text{Var}(X_t | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) &= \sum_{i=1}^p \phi_i \left(\frac{\exp(\mathbf{Z}_t^\top \delta_1)}{(1 + \exp(\mathbf{Z}_t^\top \delta_1))^2} X_{t-i} + \frac{\exp(2\mathbf{Z}_t^\top \delta_1)}{(1 + \exp(\mathbf{Z}_t^\top \delta_1))^2} X_{t-i}^2 + \frac{\exp(\mathbf{Z}_t^\top \delta_2)}{(1 + \exp(\mathbf{Z}_t^\top \delta_2))^2} (N - X_{t-i}) \right. \\ &+ \frac{\exp(2\mathbf{Z}_t^\top \delta_1)}{(1 + \exp(\mathbf{Z}_t^\top \delta_1))^2} (N - X_{t-i})^2 + 2 \prod_{i=1}^2 \frac{\exp(\mathbf{Z}_t^\top \delta_i)}{1 + \exp(\mathbf{Z}_t^\top \delta_i)} X_{t-i} (N - X_{t-i}) \left. \right) \\ &- \left(\sum_{i=1}^p \phi_i \left(\frac{\exp(\mathbf{Z}_t^\top \delta_1)}{1 + \exp(\mathbf{Z}_t^\top \delta_1)} X_{t-i} + \frac{\exp(\mathbf{Z}_t^\top \delta_2)}{1 + \exp(\mathbf{Z}_t^\top \delta_2)} (N - X_{t-i}) \right) \right)^2. \end{aligned}$$

For the detailed derivations, please see "Appendix A". Moreover, one may also interest in the autocovariance function of the RCMBAR(p)-X process. However, the derivation is complex even in the case of constant coefficients in Weiß (2009b). In this study, we rewrite the RCMBAR(p)-X process in a multivariable form, and further derive the autocovariance function. The details are given in "Appendix C".

In the following proposition, we state the strict stationary and ergodic properties of the RCMBAR(p)-X process.

Proposition 2.1 *Let $\{X_t\}_{t \in \mathbb{Z}}$ be the process defined in (2.2). If the explanatory variable sequences $\{Z_{j,t}\}$ ($j = 1, 2, \dots, q$) are all stationary sequences, then $\{X_t\}_{t \in \mathbb{Z}}$ is an irreducible, aperiodic and positive recurrent (and hence ergodic) Markov chain on state space $\mathbb{S} := \{0, 1, \dots, N\}$. Furthermore, there exists a strictly stationary process satisfying (2.2).*

The proof of Proposition 2.1 is given in "Appendix B".

3 Parameters estimation

In this section, we consider the parameter estimation problem based on a series of realizations $\{X_t\}_{t=1}^n$ from the RCMBAR(p)-X process, $\{\mathbf{Z}_t\}_{t=1}^n$ are the corresponding covariates. Denote by $\theta := (\delta_1^\top, \delta_2^\top, \phi^\top)^\top$ the parameter of interest, where $\phi = (\phi_1, \dots, \phi_{p-1})^\top$. The parameter vector takes values in the following parameter space

$$\Theta := \{ \theta \in \mathbb{R}^{q+1} \times \mathbb{R}^{q+1} \times (0, 1)^{p-1} \}.$$

In the following, we study the conditional least squares (CLS) and conditional maximum likelihood (CML) estimation methods for θ .

3.1 CLS estimation for θ

Let

$$Q(\theta) = \sum_{t=1}^n (X_t - g(\theta, X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t))^2 = \sum_{t=1}^n U_t(\theta), \tag{3.7}$$

be the CLS criterion function, where $g(\theta, X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) := E(X_t | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t)$, and $U_t(\theta) = (X_t - g(\theta, X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t))^2$. Then, the CLS-estimator $\hat{\theta}_{CLS} := (\hat{\delta}_{1,CLS}^\top, \hat{\delta}_{2,CLS}^\top, \hat{\phi}_{CLS}^\top)^\top$ is obtained by minimizing (3.7) with respect to θ , and giving

$$\hat{\theta}_{CLS} := \arg \min_{\theta \in \Theta} Q(\theta). \tag{3.8}$$

Since the RCMBAR(p)-X process is stationary and ergodic by Proposition 2.1, it follows by Theorems 3.1 and 3.2 in Klimko and Nelson (1978) that the CLS-estimators $\hat{\theta}_{CLS}$ are strongly consistent and asymptotically normally distributed. We state this property in the following theorem. The proof of this theorem is postponed to Appendix B.

Theorem 3.1 *Under the conditions of Proposition 2.1 and $E\|\mathbf{Z}_t\|^3 < \infty$, the CLS-estimators $\hat{\theta}_{CLS}$ are strongly consistent and asymptotically normal,*

$$\sqrt{n}(\hat{\theta}_{CLS} - \theta_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}^{-1} \mathbf{W} \mathbf{V}^{-1}), \tag{3.9}$$

where θ_0 is the true value of θ , $\mathbf{v} := E_{\theta_0} \left(\frac{\partial}{\partial \theta} g(\theta, X_0, \dots, X_{1-p}, \mathbf{Z}_1) \frac{\partial}{\partial \theta^\top} g(\theta, X_0, \dots, X_{1-p}, \mathbf{Z}_1) \right)$, $\mathbf{W} := E_{\theta_0} \left(\frac{\partial}{\partial \theta} g(\theta, X_0, \dots, X_{1-p}, \mathbf{Z}_1) \frac{\partial}{\partial \theta^\top} g(\theta, X_0, \dots, X_{1-p}, \mathbf{Z}_1) U_1(\theta) \right)$.

3.2 CML estimation for θ

In this section, we consider the CML estimation for θ . To this end, we need to derive the conditional likelihood function first. For fixed values of x_0, x_{-1}, \dots , and x_{1-p} , the conditional likelihood function of RCMBAR(p)-X process can be written as

$$L(\theta) = \prod_{t=1}^n P(X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}, \mathbf{Z}_t).$$

Thus, the CML-estimator $\hat{\theta}_{CML}$ can be obtained by minimizing the following conditional log likelihood function

$$\ell(\theta) = \log L(\theta) = \sum_{t=1}^n \log P(X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}, \mathbf{Z}_t),$$

and giving

$$\hat{\theta}_{CML} = \arg \max \ell(\theta). \tag{3.10}$$

The existence of (2.5) in the conditional expectations and the conditional probabilities makes the calculations of (3.8) and (3.10) very complex. It is technically very difficult or even impossible to find closed-form expressions for CLS and CML estimators. Therefore, numerical procedures have to be employed. Fortunately, we can use computer programs to complete the optimization process.

The following results establish the strong consistency and the asymptotic normality of the CML-estimators.

Theorem 3.2 *Under the conditions of Proposition 2.1, the CML-estimators $\hat{\theta}_{CML}$ are strongly consistent and asymptotically normal,*

$$\sqrt{n}(\hat{\theta}_{CML} - \theta_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}^{-1}(\theta_0)), \tag{3.11}$$

where θ_0 is the true value of θ , $\mathbf{I}(\theta)$ denotes the Fisher information matrix.

The proof of this theorem is given in "Appendix B".

4 Testing the existence of explanatory variables

In this section, we focus on an interesting issue, that is, to test whether the explanatory variables exist in the RCMBAR(p)-X model. For this purpose, we give the null hypothesis and the alternative hypothesis as follows:

$$\mathcal{H}_0 : \delta_{ij} = 0 \ (i = 1, 2, j = 1, \dots, q), \text{ v.s. } \mathcal{H}_1 : \text{At least one } \delta_{ij} \neq 0 \ (i = 1, 2, j = 1, \dots, q). \tag{4.12}$$

The inference problem in (4.12) is indeed very important as it is testing a BAR(p) model against a RCMBAR(p)-X model. When the null hypothesis holds, the model will reduce to the BAR(p) model in Weiß (2009b).

Testing problem (4.12) is equivalent to the following hypothesis:

$$\mathcal{H}_0 : \mathbf{D}\zeta = \mathbf{0} \ \text{vs} \ \mathcal{H}_1 : \mathbf{D}\zeta \neq \mathbf{0}, \tag{4.13}$$

where $\zeta = (\delta_1^\top, \delta_2^\top)^\top$, $D = \begin{pmatrix} B & \mathbf{0} \\ \mathbf{0} & B \end{pmatrix}$ is a block matrix with $B = (\mathbf{0}_{q \times 1}, I_{q \times q})$, $I_{q \times q}$ stands for a q th-order identity matrix. To address this testing problem, we develop a Wald-type test. For this purpose, we introduce some regularity conditions:

(C1) $\{X_t\}$ is a stationary process.

(C1) $\hat{\zeta} := (\hat{\delta}_1^\top, \hat{\delta}_2^\top)^\top$ is a consistent estimator of ζ . Moreover, $\hat{\zeta}$ is asymptotically normally distributed around the true value ζ_0 , i.e.,

$$\sqrt{n}(\hat{\zeta} - \zeta_0) \xrightarrow{L} N(\mathbf{0}, \Sigma),$$

for some covariance matrix Σ .

Thus, we obtain the following theorem.

Theorem 4.3 *Under the assumptions (C1–C2), the statistic for testing problem (4.13) is*

$$S_n = n\hat{\zeta}^\top D^\top (D\hat{\Sigma}D^\top)^{-1} D\hat{\zeta},$$

where $\hat{\Sigma}$ is a consistent estimator of Σ . Furthermore, when \mathcal{H}_0 is true,

$$S_n \xrightarrow{L} \chi_{2q}^2, \quad n \rightarrow \infty,$$

where χ_{2q}^2 stands for a chi-square distribution with $2q$ degrees of freedom.

Theorem 4.3 follows easily by the properties of normal distribution and the Slutsky’s Theorem. Therefore, we omit the proof of it. We can use Theorem 4.3 to test whether the autoregressive coefficient of a RCMBAR(p)-X model is a constant. Also, it can be used to test whether a specific explanatory variable is included in the model. In this point of view, it provides a way to separate the proposed model from a consistent coefficient one. In practice, the estimator $\hat{\zeta}$ can be any consistent estimator of ζ . In this study, we use the CML-estimator obtained in the previous section.

5 Forecasting for RCMBAR(p)-X process

In the following, we address the forecasting problem for the RCMBAR(p)-X process. A general method in time series forecasting is to use the conditional expectation, which yields forecasts with minimum the mean square error. However, this method is unsatisfactory for integer-valued time series, since it seldom produces integer-valued forecasts. An alternative way is to use the k -step-ahead conditional distribution (Freeland and McCabe 2004b). Provided that the k -step-ahead conditional distribution is available, point prediction such as the conditional expectation or conditional median results are easy to calculate. Yang et al. (2021) generalized (Freeland and McCabe 2004b)’s approach to a covariate-driven threshold INAR model.

In this study, we mainly focus on the one-step forecast, since it is usually often adopted in practice. A general version of k -step forecast can be easily obtained using (Freeland and McCabe 2004b)'s approach based on the representation of the RCMBAR(p)-X process given in (B.5). Notice that the state space of RCMBAR(p)-X process is a finite set on $\{0, 1, \dots, N\}$, we can easily obtain the one-step forecasting conditional distribution with parameter θ , as follows:

$$p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta) := P(X_{n+1} = x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta), \quad x = 0, 1, \dots, N, \tag{5.14}$$

where $P(X_{n+1} = x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta)$ is defined in (2.6). Based on (5.14), we can calculate the point predictions such as the conditional expectation, conditional median, and so on.

In addition to point predictions, we are also interested in the forecasting confidence interval for each point in $\{0, 1, \dots, N\}$. Given that we have already obtained some versions of $\hat{\theta}$, together with the asymptotic normality of $\hat{\theta}$ as

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(\mathbf{0}, \Sigma), \tag{5.15}$$

where θ_0 denotes the true value of θ , Σ is the covariance matrix. Then, we have the following theorem similar to Theorem 2 in Freeland and McCabe (2004b) which can be used to construct the confidence interval for $p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta)$. Obviously, the interval may be truncated outside $[0, 1]$.

Theorem 5.4 *For a fixed $x \in \{0, 1, \dots, N\}$, if assumption (5.15) holds, the quantity $p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \hat{\theta})$ has an asymptotically normal distribution with mean $p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta_0)$ and variance $n^{-1}D\Sigma D^T$, i.e.,*

$$\sqrt{n}(p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \hat{\theta}) - p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta_0)) \xrightarrow{L} N(\mathbf{0}, D\Sigma D^T),$$

where $D = \left(\frac{\partial p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right)$, $\hat{\theta}$ is the consistent estimator of θ .

The above Theorem 5.4 follows easily by (5.15) and the well-known delta method (see, e.g., van der Vaart (1998), Chapter 3). In practice, $\hat{\theta}$ can be chosen as the CML-estimator $\hat{\theta}_{CML}$ discussed in Sect. 3, and then Σ be $I^{-1}(\theta)$ accordingly. Moreover, based on Theorem 5.4, we can get the $100(1 - \alpha)$ confidence interval for $p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \theta)$ as follows:

$$C_{\theta}^{\alpha} = \left(p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \hat{\theta}) - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}, p(x|X_n, \dots, X_{n-p+1}, \mathbf{Z}_n, \hat{\theta}) + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}} \right),$$

where $\sigma = \sqrt{D\Sigma D^T}$, $u_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ -upper quantile of $N(0, 1)$.

As an illustration, we draw the one-step forecasting distribution and 95% forecasting confidence intervals under a RCMBAR(2)-X model in Fig. 1. The parameters are chosen the same as Scenario A in Sect. 6, i.e., $(\delta_{1,0}, \delta_{1,1}, \delta_{2,0}, \delta_{2,1}, \phi_1) = (0.2, 0.4, 0.4, 0.3, 0.8)$ and $N = 50$. In order to make the figure reproducible, we

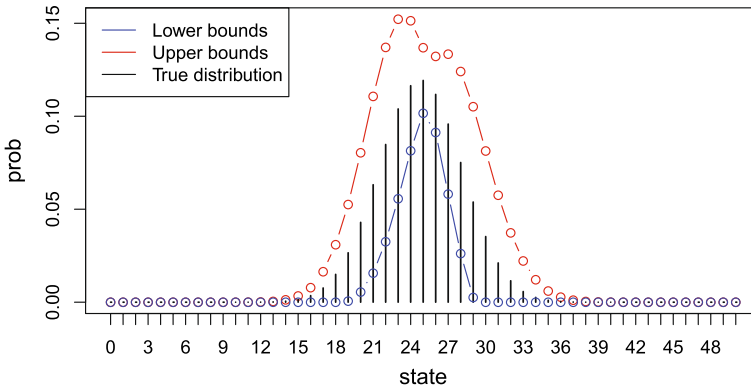


Fig. 2 One-step ahead forecasting distribution and the 95% forecasting confidence intervals

use R-code ‘set.seed(18)’ to fixed the random number. Then, we generate 200 ‘random observations’, where $X_{200} = 25$ and $X_{199} = 24$.

Figure 1 shows us that the forecasting distribution is an unimodal asymmetric distribution. The main probability points are concentrated between 12 and 40. The forecasting interval covers each probability mass. Meanwhile, the interval lengths in the middle part are greater than that on both sides. Figure 1 shows us more comprehensive statistical information about the next prediction, which is clearly more informative than a single point.

6 Simulation studies

6.1 Comparison of CLS and CML

In this subsection, we conduct simulation studies to report the performances of the proposed CLS and CML estimators. For this purpose, we choose the sample sizes $n = 100, 300$ and 500 for the following two models:

Scenario A. In this scenario, we consider a RCMBAR(2)-X model with parameters $(\delta_{1,0}, \delta_{1,1}, \delta_{2,0}, \delta_{2,1}, \phi_1) = (0.2, 0.4, 0.4, 0.3, 0.8)$ and $N = 50$. The explanatory variable $Z_{1,t}$ is generated from an i.i.d. $N(0, 1)$ distribution.

Scenario B. In this scenario, we consider a RCMBAR(3)-X model with parameters $(\delta_{1,0}, \delta_{1,1}, \delta_{2,0}, \delta_{2,1}, \delta_{2,2}, \phi_1, \phi_2) = (0.4, 0.1, 0.2, 0.6, 0.4, 0.3)$ and $N = 50$. The explanatory variable is generated from an AR(1) process, $Z_{1,t} = 0.2Z_{1,t-1} + \epsilon_t$ with $\epsilon_t \sim N(0, 1)$ and $Z_{1,0} = 0$.

Scenario C. In this scenario, we consider a RCMBAR(2)-X model with parameters $(\delta_{1,0}, \delta_{1,1}, \delta_{1,2}, \delta_{2,0}, \delta_{2,1}, \delta_{2,2}, \phi_1) = (0.2, 0.4, 0.6, 0.1, 0.3, 0.5, 0.7)$ and $N = 40$. There are two explanatory variables in the model, where $Z_{1,t}$ is generated from an i.i.d. $N(0, 1)$ distribution, $Z_{2,t}$ is generated from an AR(1) process, $Z_{2,t} = 0.5Z_{2,t-1} + \epsilon_t$ with $\epsilon_t \sim N(0, 1)$ and $Z_{2,0} = 0$.

The above three scenarios consider different cases of explanatory variables. Scenarios A considers a simple independent normally distributed explanatory variable. Scenarios B considers a dependence explanatory variable of an AR(1) process. In Scenarios C, we consider two explanatory variables in the model. Firstly, we show the sample paths and autocorrelation function (ACF) plots for the two scenarios in Fig. 2. As is seen in Fig. 2 that there is no trend and seasonal characteristics in the subfigures, indicating that all series are stationary. Moreover, the three series show different autocorrelation characteristics, which implies that RCMBAR(p)-X model can describe different autocorrelation structures.

Next, we conduct simulation studies to show the performances of the proposed CLS and CML estimators. For the above three models, we calculated the estimates based on the two methods, the empirical biases (Bias), and the mean square errors (MSE). All the simulations are performed under the \mathcal{R} software based on 1000 replications. The simulation results are summarized in Tables 1, 2 and 3.

It can be seen from Tables 1, 2 and 3 that the biases and MSEs are getting small with the increases of the sample size, indicating the consistency of the estimators. Generally, the CML estimates seem to be more efficient since they present smaller bias and MSE values, regardless the number and the type of explanatory variables.

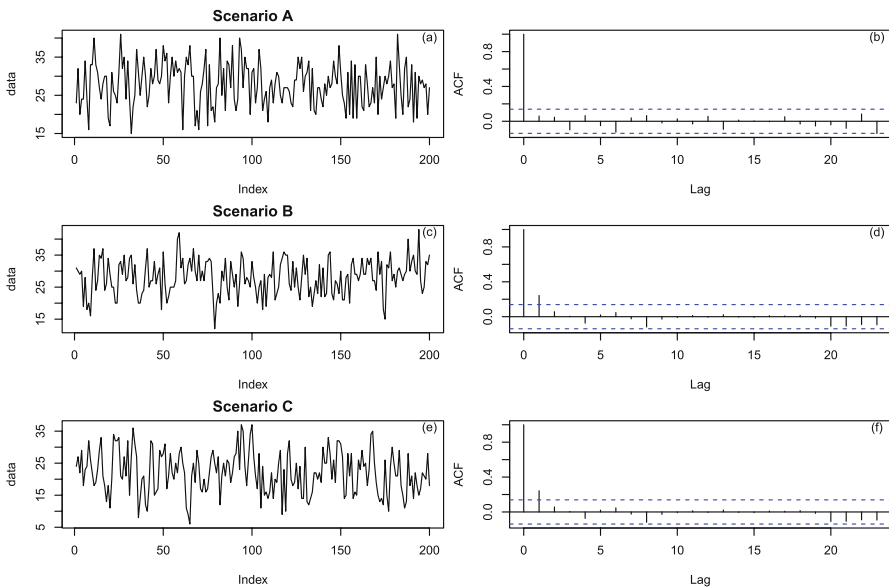


Fig. 3 Sample path and ACF plots of Scenarios A, B and C

Table 1 Simulation results of Scenario A under different sample sizes

Methods	n	Results	$\delta_{1,0} = 0.2$	$\delta_{1,1} = 0.4$	$\delta_{2,0} = 0.4$	$\delta_{2,1} = 0.3$	$\phi_1 = 0.8$
CLS	100	Bias	0.0165	0.0159	0.0187	0.0155	0.0647
		MSE	0.0004	0.0004	0.0004	0.0001	0.0030
	300	Bias	0.0135	0.0150	0.0142	0.0078	0.0344
		MSE	0.0001	0.0002	0.0003	0.0001	0.0010
	500	Bias	0.0056	0.0072	0.0123	0.0053	0.0141
		MSE	0.0001	0.0001	0.0002	0.0000	0.0003
CML	100	Bias	0.0119	-0.0201	0.0139	-0.0060	-0.0237
		MSE	0.0001	0.0004	0.0003	0.0001	0.0006
	300	Bias	0.0035	-0.0109	-0.0107	0.0089	0.0183
		MSE	0.0001	0.0002	0.0002	0.0001	0.0004
	500	Bias	0.0037	0.0047	-0.0044	-0.0001	0.0099
		MSE	0.0000	0.0000	0.0000	0.0000	0.0001

Table 2 Simulation results of Scenario B under different sample sizes

Methods	n	Results	$\delta_{1,0} = 0.4$	$\delta_{1,1} = 0.1$	$\delta_{2,0} = 0.2$	$\delta_{2,1} = 0.6$	$\phi_1 = 0.4$	$\phi_2 = 0.3$
CLS	100	Bias	-0.0223	0.0261	0.0208	0.0124	0.0295	0.0310
		MSE	0.0006	0.0007	0.0006	0.0002	0.0007	0.0007
	300	Bias	-0.0294	0.0168	0.0202	0.0197	0.0160	0.0116
		MSE	0.0104	0.0003	0.0004	0.0003	0.0002	0.0002
	500	Bias	-0.0065	0.0144	0.0119	0.0108	0.0118	0.0096
		MSE	0.0001	0.0002	0.0002	0.0001	0.0001	0.0001
CML	100	Bias	-0.0120	0.0222	0.0248	-0.0126	0.0212	-0.0262
		MSE	0.0001	0.0005	0.0006	0.0002	0.0006	0.0007
	300	Bias	-0.0082	0.0198	0.0186	-0.0133	0.0161	-0.0092
		MSE	0.0001	0.0004	0.0003	0.0003	0.0003	0.0001
	500	Bias	-0.0059	0.0114	0.0025	-0.0020	0.0120	-0.0031
		MSE	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000

6.2 Powers of the test

In this subsection, we conduct simulations to show the performances of the hypothesis test discussed in Sect. 4. To this end, we further consider the following two Scenarios:

Scenario D. In this scenario, we also consider a RCMBAR(2)-X model with parameters $(\delta_{1,0}, \delta_{1,1}, \delta_{2,0}, \delta_{2,1}, \phi_1) = (0.2, 0, 0.4, 0, 0.8)$ and $N = 40$. The explanatory variable $Z_{1,t}$ is generated in the same way as Scenario A.

Table 3 Simulation results of Scenario C under different sample sizes

Methods	n	Results	$\delta_{1,0} = 0.2$	$\delta_{1,1} = 0.4$	$\delta_{1,2} = 0.6$	$\delta_{2,0} = 0.1$	$\delta_{2,1} = 0.3$	$\delta_{2,2} = 0.5$	$\phi_1 = 0.7$
CLS	100	Bias	-0.0096	0.0025	0.0114	0.0154	0.0081	0.0096	0.1074
		MSE	0.0164	0.0172	0.0156	0.0189	0.0183	0.0162	0.5885
	300	Bias	-0.0071	-0.0015	0.0020	0.0098	0.0047	0.0043	0.0689
		MSE	0.0059	0.0047	0.0048	0.0071	0.0055	0.0052	0.3851
	500	Bias	-0.0043	-0.0011	-0.0004	0.0080	0.0031	0.0038	0.0614
		MSE	0.0033	0.0028	0.0025	0.0040	0.0032	0.0027	0.2747
CML	100	Bias	-0.0130	-0.0007	0.0057	0.0170	0.0098	0.0169	0.1772
		MSE	0.0002	0.0000	0.0000	0.0003	0.0001	0.0003	0.0314
	300	Bias	-0.0030	-0.0011	-0.0019	0.0060	0.0056	0.0082	0.0768
		MSE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0059
	500	Bias	-0.0035	-0.0003	0.0004	0.0048	0.0036	0.0040	0.0850
		MSE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0072

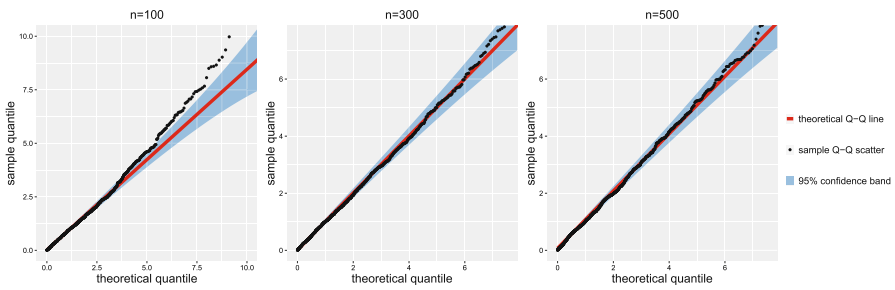


Fig. 4 Q-Q plots of S_n under Scenario D based on CML method

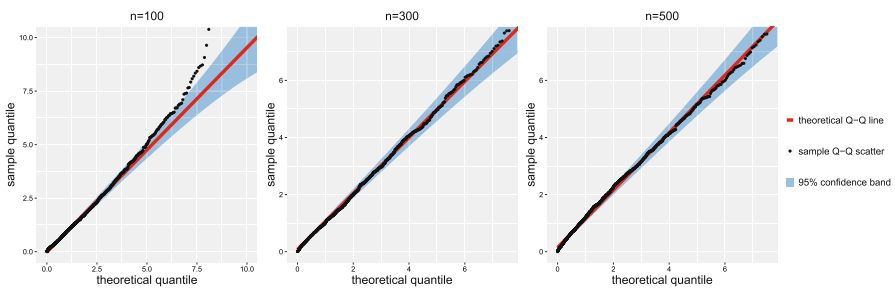


Fig. 5 Q-Q plots of S_n under Scenario E based on CML method

Scenario E. In this scenario, we consider a RCMBAR(3)-X model with parameters $(\delta_{1,0}, \delta_{1,1}, \delta_{2,0}, \delta_{2,1}, \phi_1, \phi_2) = (0.6, 0, 0.3, 0, 0.5, 0.3)$ and $N = 40$. The explanatory variable $Z_{1,t}$ is generated in the same way as Scenario B.

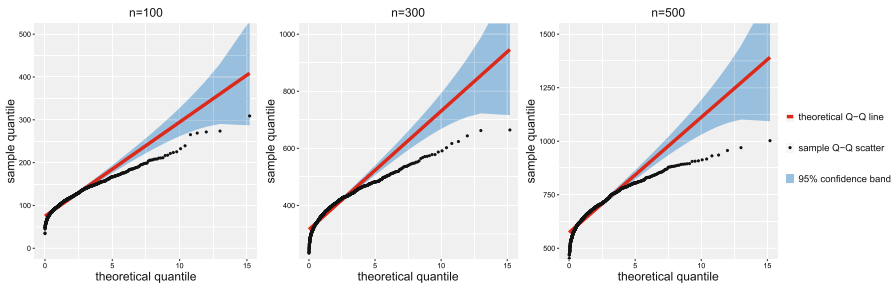


Fig. 6 Q-Q plots of S_n under Scenario A based on CML method

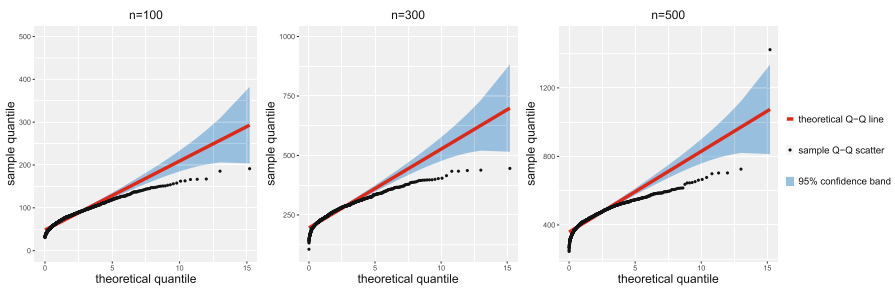


Fig. 7 Q-Q plots of S_n under Scenario B based on CML method

Table 4 Empirical power and size of test (4.13) based on CML method

n	Empirical power when \mathcal{H}_0 is not true		Empirical size when \mathcal{H}_0 is true	
	Scenario A	Scenario B	Scenario D	Scenario E
100	1	1	0.045	0.055
300	1	1	0.047	0.051
500	1	1	0.051	0.050

It is clear that Scenarios D and E are cases where \mathcal{H}_0 is true. Firstly, we give an intuitive explanation for Theorem 4.3. For this purpose, we draw the Q-Q plots of S_n under Scenarios D and E in Figs. 3 and 4, aiming to show how S_n distributes when \mathcal{H}_0 is true. Meanwhile, we also draw the Q-Q plots of S_n under Scenarios A and B in Figs. 5 and 6, aiming to investigate whether the result of chi-square distribution for S_n will still hold when \mathcal{H}_0 is not true.

As is seen in Figs. 3 and 4 that, the sample Q-Q scatter plots getting closer to the theoretical Q-Q lines as the sample size increases. This implies the testing statistics S_n gradually converges to a χ^2_2 distribution as expected, regardless of the order of the model. On the contrary, as is seen in Figs. 5 and 6 that the

scatter plots all fall outside the confidence band areas, indicating that the χ_2^2 distribution is no longer valid.

Next, we show the detailed performances of testing problem (4.13) discussed in Section 4. To this end, we summary the simulation results under Scenarios A, B, D, and E using CML method in Table 4. As is seen in Table 4 that when \mathcal{H}_0 is true (Scenarios D and E), the empirical size is getting closer to the significance level of 0.05, which implies that the asymptotic distribution in Theorem 4.3 is correct. On the other hand, we also see that all empirical power results (Scenarios A and B) are equal to one when \mathcal{H}_0 is not true. This implies the proposed test statistics performs well in practice.

7 Real data example

In this section, we will use the RCMBAR(p)-X model to fit a set of rainy days at Bremen in Germany. The data was published by the German Weather Service, and can be downloaded in the following URL: <http://www.dwd.de/>. The original data set records the local daily rainfall of Bremen. We choose the time period from January 2011 to December 2021. With the selected data set, we calculated the number of rainfall days per week and the corresponding rainfall. Specifically, for each week t , the value X_t counts the number of rainy days, and $Z_{1,t}$ records the total rainfall of the week. Therefore, we obtain a time series of counts with a finite-range $N = 7$, totally consists 574 weekly observations. Moreover, in this study, $Z_{1,t}$ is used as an explanatory variable.

For convenience, we denote $\{X_t\}_{t=0}^{573}$ and $\{Z_{1,t}\}_{t=0}^{573}$ as the sequences of observed data and explanatory variables, and further draw the time series and ACF plots of the observations in Fig. 7, draw the time series of the covariate in 8. From Fig. 7

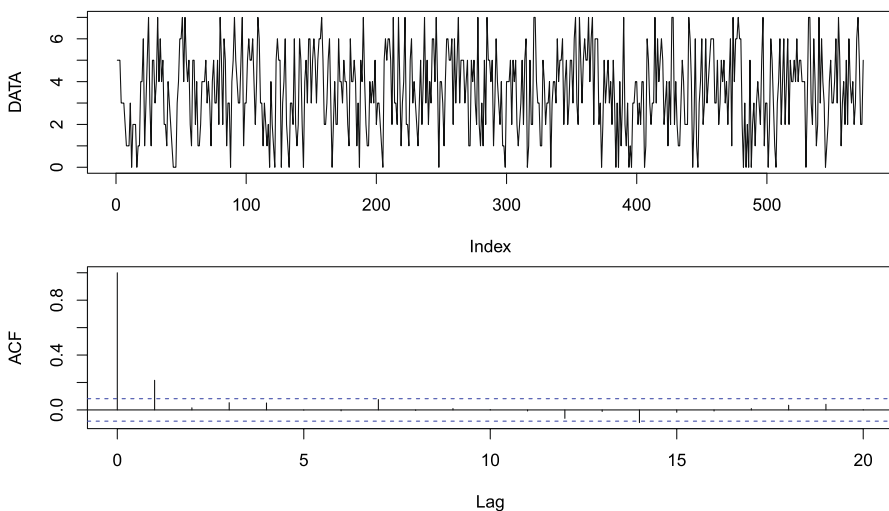


Fig. 8 Time series and ACF plots of the rainfall days counts

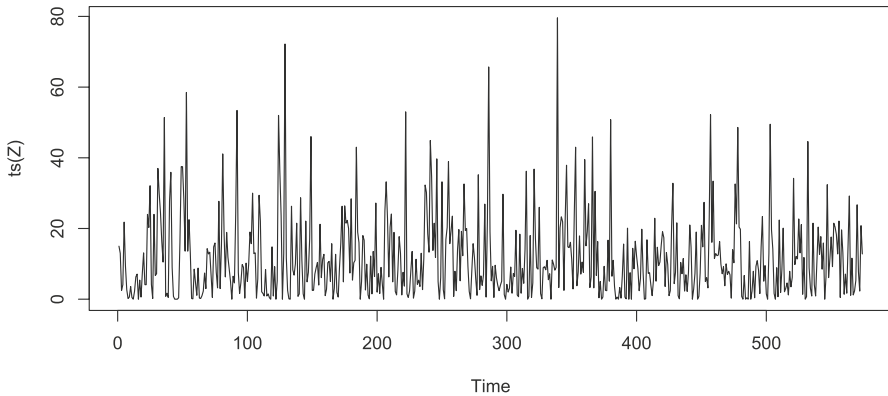


Fig. 9 Time series plot of the covariates

we can see that the analyzed data set is a stationary time series. The ACF exhibits an exponential decay trend. Figure 8 also implies the sequence of covariate is stationary.

For comparison purpose, we also use the BAR(1) model (McKenzie 1985), the BAR(p) model (Weiß 2009b) with $p = 2$ and 3, and the CDBAR(1) model (Wang et al. 2021) to fit this data set, and compare different models by AIC and BIC criteria. The BAR(1) model is the original binomial autoregressive model of order one, which does not contain explanatory variables. The BAR(p) model is an extension of BAR(1) model, which is a p th-order constant coefficients binomial autoregressive model with the j th-order regime existing in the model with probability ϕ_j ($j = 1, 2, \dots, p$). The CDBAR(1) model is defined via introducing explanatory variables into both autoregressive coefficients of a BAR(1) model, which is also a special case of the RCMBAR(p)-X model proposed in this study. For each of the fitted model, we calculate the conditional maximum likelihood estimation (CMLE) of the model parameters, the corresponding standard errors (SE), AIC and BIC values. All the fitting results are summarized in Table 5.

It can be seen from Table 4 that (i) among similar models, higher order models have better fitting effect than lower order models; and (ii) among different models, the models with explanatory variables are better than the models without explanatory variables. This shows that it is necessary to study high-order models and consider explanatory variables. Moreover, among all competition models, the RCMBAR(3)-X model has the smallest AIC and BIC values. This implies that the RCMBAR(3)-X model is a competitive model in terms of AIC and BIC, and is appropriate for fitting this data set.

In the following, we conduct the diagnostic checking for the fitted RCMBAR(3)-X model. For this purpose, we need to calculate the standardized Person residuals. As is reviewed by many authors (see, e.g., Yang et al. (2022, 2023), Zhang et al.

Table 5 Fitting results of the rainfall days counts under different models

Models	Para	SE	AIC	BIC
BAR (1)	$\hat{\rho} = 0.1486$	0.0210	2525.030	2533.732
	$\hat{\pi} = 0.5103$	0.0074		
BAR (2)	$\hat{\rho} = 0.1636$	0.0255	2523.376	2536.418
	$\hat{\pi} = 0.5099$	0.0066		
	$\hat{\phi}_1 = 0.9096$	0.1074		
BAR (3)	$\hat{\rho} = 0.2386$	0.0301	2513.742	2531.125
	$\hat{\pi} = 0.5101$	0.0073		
	$\hat{\phi}_1 = 0.6571$	0.0860		
	$\hat{\phi}_2 = 0.0667$	0.0776		
CDBAR (1)	$\hat{\theta}_1 = 0.0412$	0.0038	2307.994	2316.696
	$\hat{\lambda}_1 = 0.0215$	0.0038		
RCMBAR (2)-X	$\hat{\delta}_{1,0} = -0.9454$	0.1325	2071.353	2093.098
	$\hat{\delta}_{1,1} = 0.1376$	0.0104		
	$\hat{\delta}_{2,0} = -0.7097$	0.0881		
	$\hat{\delta}_{2,1} = 0.0345$	0.0040		
	$\hat{\phi}_1 = 0.6083$	0.0779		
RCMBAR (3)-X	$\hat{\delta}_{1,0} = -1.0539$	0.1491	2041.910	2067.994
	$\hat{\delta}_{1,1} = 0.1737$	0.0125		
	$\hat{\delta}_{2,0} = -0.7796$	0.1030		
	$\hat{\delta}_{2,1} = 0.0295$	0.0045		
	$\hat{\phi}_1 = 0.5000$	0.0718		
	$\hat{\phi}_2 = 0.1838$	0.0612		

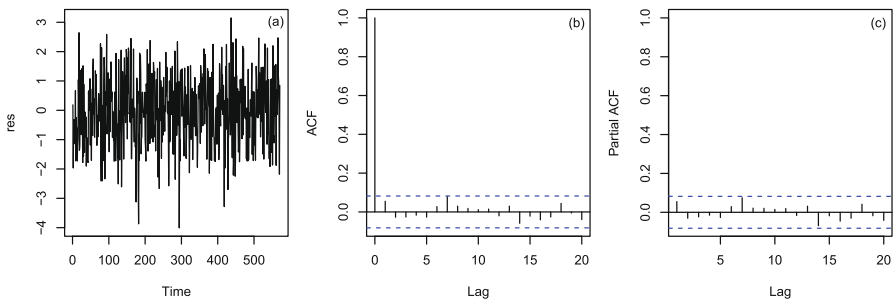


Fig. 10 Diagnostic checking plots in fitting RCMBAR(3)-X model with the rainfall data set. **a** standardized residuals; **b** ACF plot of the residuals; **c** PACF plot of the residuals

(2022) and among others), the standardized residuals provides a relatively easy way to check whether the model fits data adequately. Specifically, if the model is

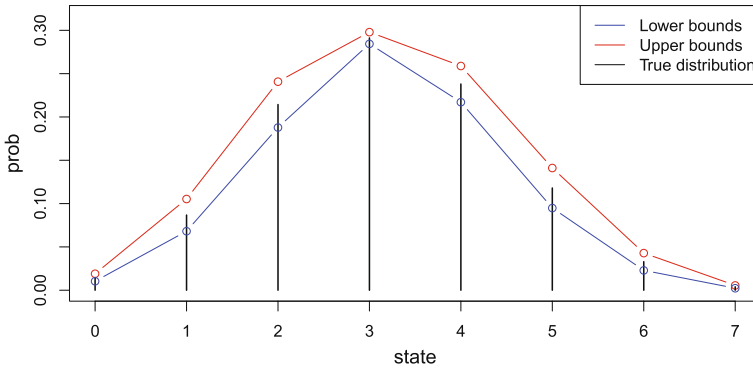


Fig. 11 Forecasting distribution and the 95% confidence intervals of the analyzed data sets

correctly specified, the residuals should have no significant serial correlation. For the RCMBAR(3)-X model, the standardized residuals is defined as

$$e_t = \frac{X_t - E(X_t|X_{t-1}, \dots, X_{t-3}, Z_{1,t})}{\sqrt{\text{Var}(X_t|X_{t-1}, \dots, X_{t-3}, Z_{1,t})}}, \quad t = 1, 2, \dots, n. \tag{7.16}$$

In practice, we can substitute the CMLE results into the conditional expectation and conditional variance equations in (7.16) to calculate $\{\hat{e}_t\}$.

Figure 9 shows the time series plot, ACF and partial autocorrelation function (PACF) plots of the the standardized residuals under RCMBAR(3)-X model. As is shown in Fig. 9 that the residuals is a stationary series. The p -value of ADF test is smaller than 0.01, which ensures the stationarity of the residuals. Moreover, the ACF and PACF plots show that the residuals have no sequence autocorrelation. This implies $\{\hat{e}_t\}$ is a stationary white noise which ensures that the RCMBAR(3)-X model is correctly specified.

Finally, as an application, we draw the one-step ahead forecasting distribution and the forecasting confidence intervals of the corresponding points in Fig. 10. From Fig. 10 we can see that the most likely number of rainy days in the region next week are 3 - 4 days.

8 Conclusions

This article introduces a p th-order random coefficients integer-valued binomial autoregressive process with explanatory variables, which can accurately capture the higher-order dependence of integer-valued time series with bounded support, and conveniently model the relationship between observational process with covariates. The CLS and CML methods are introduced to address the parameter estimation problems for the model. The results show that the CML method has higher estimation accuracy. Moreover, we also considered the existence test of explanatory variables. Finally, a real data example is provided to show the outstanding performance of the proposed model.

Appendix A: The derivations of moments

In the following, we derive the derivations of moments for the RCMBAR(p)-X model. With the representation of (2.4), the calculation of conditional expectation is trivial, thereby, omitted. We go on to derive the conditional variance.

Notice that (2.4) implies $D_{t,i}D_{t,j} = 0$ for $i \neq j$, it follows that

$$\begin{aligned}
 & E(X_t^2 | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) \\
 &= E\left(\sum_{i=1}^p D_{t,i}^2 (\alpha_t \circ X_{t-i} + \beta_t \circ (N - X_{t-i}))^2 | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t\right) \\
 &= \sum_{i=1}^p \phi_i E[(\alpha_t \circ X_{t-i})^2 + 2(\alpha_t \circ X_{t-i})(\beta_t \circ (N - X_{t-i})) + (\beta_t \circ (N - X_{t-i}))^2 | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t] \\
 &= \sum_{i=1}^p \phi_i [\alpha_i(1 - \alpha_i)X_{t-i} + \alpha_i^2 X_{t-i}^2 + \beta_i(1 - \beta_i)(N - X_{t-i}) + \beta_i^2(N - X_{t-i})^2 + 2\alpha_i\beta_i X_{t-i}(N - X_{t-i})] \\
 &= \sum_{i=1}^p \phi_i \left(\frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^2} X_{t-i} + \frac{\exp(2\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^2} X_{t-i}^2 + \frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2))^2} (N - X_{t-i}) \right. \\
 &\quad \left. + \frac{\exp(2\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^2} (N - X_{t-i})^2 + 2 \prod_{i=1}^2 \frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_i)}{1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_i)} X_{t-i}(N - X_{t-i}) \right).
 \end{aligned}$$

Therefore, the conditional variance can be derived as follows

$$\begin{aligned}
 & \text{Var}(X_t | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) = E(X_t^2 | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) - E^2(X_t | X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) \\
 &= \sum_{i=1}^p \phi_i \left(\frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^2} X_{t-i} + \frac{\exp(2\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^2} X_{t-i}^2 + \frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2))^2} (N - X_{t-i}) \right. \\
 &\quad \left. + \frac{\exp(2\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^2} (N - X_{t-i})^2 + 2 \prod_{i=1}^2 \frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_i)}{1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_i)} X_{t-i}(N - X_{t-i}) \right) \\
 &\quad - \left(\sum_{i=1}^p \phi_i \left(\frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1)} X_{t-i} + \frac{\exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2)}{1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2)} (N - X_{t-i}) \right) \right)^2.
 \end{aligned}$$

Appendix B: The proofs of theorems

Proof of Proposition 2.1 We should first prove the RCMBAR(p)-X process defined by (2.2) is an irreducible and aperiodic Markov chain. Without loss of generality, denote by $(\Omega_j, \mathcal{A}_j, P_j)$ the probability space of $Z_{j,t}$. It follows by Definition 1 that $E|Z_{j,t}| = \int_{\Omega_j} |Z_{j,t}| dP_j < \infty$, which implies

$$I_Z(i, m, t) := \int_{\Omega_1} \dots \int_{\Omega_q} \frac{\exp(m\mathbf{Z}_t^\top \boldsymbol{\delta}_1)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_1))^{x_{t-i}}} \frac{\exp((x_t - m)\mathbf{Z}_t^\top \boldsymbol{\delta}_2)}{(1 + \exp(\mathbf{Z}_t^\top \boldsymbol{\delta}_2))^{N-x_{t-i}}} dZ_{1,t} \dots dZ_{q,t} < \infty.$$

Notice that each term in (2.5) is strictly greater than zero, we obtain

$$\begin{aligned}
 &P(X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}) \\
 &= \int_{\Omega_1} \dots \int_{\Omega_q} P(X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}, \mathbf{Z}_t) dZ_{1,t} \dots dZ_{q,t} \\
 &= \sum_{i=1}^p \phi_i \sum_{m=a}^b \binom{x_{t-i}}{m} \binom{N - x_{t-i}}{x_t - m} I_Z(i, m, t) > 0.
 \end{aligned} \tag{A.1}$$

Equation (A.1) implies the process (2.2) is an irreducible and aperiodic chain. Since the state space $\mathbb{S} := \{0, 1, \dots, N\}$ has only a finite number of elements, $\{X_t\}_{t \in \mathbb{Z}}$ is also a positive recurrent Markov chain and hence ergodic. Finally, Theorem 1.3 in Karlin and Taylor (1975) guarantees the existence of the stationary distribution for $\{X_t\}$. \square

Proof of Theorem 3.1 In order to prove Theorem 3.1, we need to check all the regularity conditions of Theorems 3.1 and 3.2 in Klimko and Nelson (1978) hold. The regularity conditions for Theorem 3.1 in Klimko and Nelson (1978) are given as follows:

(i) $\partial g / \partial \theta_i, \partial^2 g / \partial \theta_i \partial \theta_j, \partial^3 g / \partial \theta_i \partial \theta_j \partial \theta_k$ exist and are continuous for all $\theta \in \Theta$, where $\theta_i, \theta_j, \theta_k$ denote the components of $\theta, i, j, k \in \{1, 2, \dots, m\}$, g is the abbreviation for $g(\theta, X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t)$, $m = 2q + p + 1$ denotes the dimension of θ ;

(ii) For $i, j \in \{1, 2, \dots, m\}$, $E|(X_1 - g) \partial g / \partial \theta_i| < \infty, E|(X_1 - g) \partial^2 g / \partial \theta_i \partial \theta_j| < \infty$ and $E|\partial g / \partial \theta_i \cdot \partial g / \partial \theta_j| < \infty$, where g and its partial derivatives are evaluated at θ_0 and the σ -field generated by all the information before zero time;

(iii) For $i, j, k \in \{1, 2, \dots, m\}$ there exist functions $H^{(0)}(X_0, \dots, X_{1-p}), H_i^{(1)}(X_0, \dots, X_{1-p}), H_{ij}^{(2)}(X_0, \dots, X_{1-p}), H_{ijk}^{(3)}(X_0, \dots, X_{1-p})$ such that

$$|g| < H^{(0)}, |\partial g / \partial \theta_i| < H_i^{(1)}, |\partial^2 g / \partial \theta_i \partial \theta_j| < H_{ij}^{(2)}, |\partial^3 g / \partial \theta_i \partial \theta_j \partial \theta_k| < H_{ijk}^{(3)}, \tag{B.1}$$

for all $\theta \in \Theta$, and

$$\begin{aligned}
 &E|X_1 \cdot H_{ijk}^{(3)}(X_0, \dots, X_{1-p})| < \infty, \\
 &E(H^{(0)}(X_0, \dots, X_{1-p}) \cdot H_{ijk}^{(3)}(X_0, \dots, X_{1-p})) < \infty, \\
 &E(H_i^{(1)}(X_0, \dots, X_{1-p}) \cdot H_{ij}^{(2)}(X_0, \dots, X_{1-p})) < \infty.
 \end{aligned} \tag{B.2}$$

Recall that $g(\theta, X_{t-1}, \dots, X_{t-p}, \mathbf{Z}_t) = \sum_{i=1}^p \phi_i \left(\frac{\exp(\mathbf{Z}_t^\top \delta_1)}{1 + \exp(\mathbf{Z}_t^\top \delta_1)} X_{t-i} + \frac{\exp(\mathbf{Z}_t^\top \delta_2)}{1 + \exp(\mathbf{Z}_t^\top \delta_2)} (N - X_{t-i}) \right)$. It is easy to check that condition (i) holds. Denote by $p_k := P(X_t = k), k = 0, 1, \dots, N$. Thus, we have for any fixed $s \geq 1$,

$$E(|X_t|^s) = E(X_t^s) = \sum_{k=0}^N p_k \cdot k^s \leq \sum_{k=0}^N p_k \cdot N^s = N^s < \infty. \tag{B.3}$$

By Definition 1, \mathbf{Z}_t has a finite covariance matrix, which, together with (B.3) ensures that condition (ii) holds. Denote by $\mathbf{x} = (x_1, \dots, x_n)^\top$ a n -dimensional vector, further denote by $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \dots + |x_n|$, $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$. Let

$$\begin{aligned} H^{(0)}(X_0, \dots, X_{1-p}) &= N, & H_i^{(1)}(X_0, \dots, X_{1-p}) &= N\|\mathbf{Z}_t\|_1, \\ H_{ij}^{(2)}(X_0, \dots, X_{1-p}) &= N\|\mathbf{Z}_t\|_\infty^2, & H_{ijk}^{(3)}(X_0, \dots, X_{1-p}) &= N\|\mathbf{Z}_t\|_\infty^3, \end{aligned} \tag{B.4}$$

then for any $i, j, k \in \{1, 2, \dots, m\}$, we can verify that (B.1) holds. Moreover, $E\|\mathbf{Z}_t\|^3 < \infty$ and (B.4) imply that (B.2) holds, which implies that $\hat{\theta}_{CLS}$ is a strongly consistent estimator. With the fact that $X_t - g$ is bounded by $2N$, we obtain that $U_t(\theta)$ is bounded by $4N^2$. Together with condition (ii), we have that

$$E(U_1(\theta)|\partial g/\partial\theta_i \cdot \partial g/\partial\theta_j|) < \infty, \quad i, j \in \{1, 2, \dots, m\}.$$

Therefore, the regularity conditions for Theorem 3.2 in Klimko and Nelson (1978) hold are also hold, implying the asymptotic normality for $\hat{\theta}_{CLS}$. \square

Proof of Theorem 3.2 To prove Theorem 3.2, we first give an equivalent representation of the RCMBAR(p)- X process. We begin with some notations. Let $\mathbf{Y}_t := (X_t, X_{t-1}, \dots, X_{t-p+1})^\top$, $\alpha_{i,t} = D_{i,t}\alpha_t$, $\beta_{i,t} = D_{i,t}\beta_t$, $i = 1, 2, \dots, p$, and further denote two p th-order matrices α_t and β_t as

$$\alpha_t := \begin{pmatrix} \alpha_{1,t} & \alpha_{2,t} & \dots & \alpha_{p-1,t} & \alpha_{p,t} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \beta_t := \begin{pmatrix} \beta_{1,t} & \beta_{2,t} & \dots & \beta_{p-1,t} & \beta_{p,t} \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

With the fact that $0 \circ X = 0$ and $1 \circ X = X$ (see Lemma 1 in Silva and Oliveira 2004), model (2.2) can be written in the following form:

$$\mathbf{Y}_t = \alpha_t \circ \mathbf{Y}_{t-1} + \beta_t \circ (N - \mathbf{Y}_{t-1}), \tag{B.5}$$

where $N^\top = (N, \dots, N)_{1 \times p}$, ‘‘ \circ ’’ operation here denotes a matrix operation which acts as the usual matrix multiplication while replacing scaler multiplication with the binomial thinning operation. Thus, we obtain a multivariate version binomial autoregressive model with state space

$$\mathbb{S} := \underbrace{\{0, 1, \dots, N\} \times \{0, 1, \dots, N\} \dots \times \{0, 1, \dots, N\}}_{p \text{ multiple cartesian product}},$$

i.e., $\mathbb{S} = \{(s_1, \dots, s_p) | s_j \in \{0, 1, \dots, N\}, j = 1, 2, \dots, p\}$. Denote by $P_{t|t-1}(\theta) := P(\mathbf{Y}_t = \mathbf{y}_t | \mathbf{Y}_{t-1} = \mathbf{y}_{t-1})$ the transition probability of $\{\mathbf{Y}_t\}$. Thus, we have

$$\begin{aligned} P_{t|t-1}(\theta) &: = P(\mathbf{Y}_t = \mathbf{y}_t | \mathbf{Y}_{t-1} = \mathbf{y}_{t-1}, \mathbf{Z}_t) \\ &= P(X_t = x_t, \dots, X_{t-p+1} = x_{t-p+1} | X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}, \mathbf{Z}_t) \\ &= P(X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}, \mathbf{Z}_t). \end{aligned} \tag{B.6}$$

Equations (B.6) implies models (2.2) and (B.5) have the same transition probabilities, and also have the same CML estimators accordingly. Therefore, we can use the result in Billingsley (1961) to prove Theorem 3.2. To this end, we need to verify that Condition 5.1 in Billingsley (1961) holds. Condition 5.1 of Billingsley (1961) is fulfilled provided that:

1. The set D of (k, l) such that $P_{l|l-1}(\theta) = P(Y_t = k | Y_{t-1} = l, Z_t) > 0$ is independent of θ ;
2. Each $P_{l|l-1}(\theta)$ has continuous partial derivatives of third order throughout Θ ;
3. The $d \times r$ matrix

$$(\partial P_{l|l-1}(\theta) / \partial \theta_u), \quad u = 1, \dots, r, \tag{B.7}$$

(d being the number of elements in D) has rank r throughout Θ , $r := \dim(\Theta)$.

4. For each $\theta \in \Theta$ there is only one ergodic set and there no transient states.

Conditions 1 and 2 are easily fulfilled by (2.6). For any θ , we can select a r dimensional square matrix of rank r from the $d \times r$ dimensional matrix (B.7), then Condition 3 is also true. Since the state space of (B.5) is a finite set, and $P_{l|l-1}(\theta) > 0$, then Condition 4 holds. Thus, Conditions 1 to 4 are all fulfilled, which implies Condition 5.1 in Billingsley (1961) holds. Thereby, the CML-estimators $\hat{\theta}_{CML}$ are strongly consistent and asymptotically normal.

The proof is complete. □

Appendix C: the autocovariance function

Based on model (B.5), we derive the autocovariance function in the following form. Denote by $\gamma(k) := \text{Cov}(Y_t, Y_{t-k})$. By the law of total covariance, we have

$$\begin{aligned} \gamma(k) &= \text{Cov}(E(Y_t | Y_{t-1}, \dots), E(Y_{t-k} | Y_{t-1}, \dots)) + 0 \\ &= \text{Cov}(E(\alpha_t)Y_{t-1} + E(\beta_t)(N - Y_{t-1}), Y_{t-k}) \\ &= E(\alpha_t)\text{Cov}(Y_{t-1}, Y_{t-k}) - E(\beta_t)\text{Cov}(Y_{t-1}, Y_{t-k}) \\ &= (E(\alpha_t) - E(\beta_t))\gamma(k - 1) \\ &= \dots \\ &= (E(\alpha_t) - E(\beta_t))^k \gamma(0). \end{aligned}$$

The above representation gives a measure of autocorrelation property of (B.5), which further gives a measure for the RCMBAR(p)-X process.

Acknowledgements We gratefully acknowledge the associate editor, and anonymous referees for their valuable time and helpful comments that have helped improve this article substantially.

Funding This work is supported by National Natural Science Foundation of China (No. 11901053), Natural Science Foundation of Jilin Province (Nos. YDZJ202301ZYTS393, 20230201078GX,

20220101038JC, 20210101149JC), Postdoctoral Foundation of Jilin Province (No. 2023337), Scientific Research Project of Jilin Provincial Department of Education (Nos. JJKH2022 0671KJ, JJKH20230665KJ).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Al-Osh MA, Alzaid AA (1987) First-order integer-valued autoregressive (INAR(1)) process. *J Time Ser Anal* 8(3):261–275. <https://doi.org/10.1111/j.1467-9892.1987.tb00438.x>
- Billingsley P (1961) *Statistical inference for Markov processes*. University of Chicago Press, Chicago
- Brännäs K (1995) Explanatory variables in the AR(1) count data model. *Umeå Econ Stud* 381
- Brännäs K, Nordström J (2006) Tourist accommodation effects of festivals. *Tour Econ* 12(2):291–302. <https://doi.org/10.5367/000000006777637458>
- Chen CWS, Khamthong K, Lee S (2019) Markov switching integer-valued generalized auto-regressive conditional heteroscedastic models for dengue counts. *J R Stat Soc Ser C Appl Stat* 68(4):963–983. <https://doi.org/10.1111/rssc.12344>
- Chen H, Li Q, Zhu F (2020) Two classes of dynamic binomial integer-valued ARCH models. *Braz J Probab Stat* 34:685–711. <https://doi.org/10.1214/19-BJPS452>
- Ding X, Wang D (2016) Empirical likelihood inference for INAR(1) model with explanatory variables. *J Korean Stat Soc* 45(4):623–632. <https://doi.org/10.1016/j.jkss.2016.05.004>
- Enciso-Mora V, Neal P, Rao TS (2009) Integer valued AR processes with explanatory variables. *Sankhyā: The Indian J Stat*; 71(2):248–263. <http://www.jstor.org/stable/41343031>
- Freeland RK, McCabe BPM (2004) Analysis of low count time series data by Poisson autoregression. *J Time Ser Anal* 25(5):701–722. <https://doi.org/10.1111/j.1467-9892.2004.01885.x>
- Freeland RK, McCabe BPM (2004) Forecasting discrete valued low count time series. *Int J Forecast* 20(3):427–434. [https://doi.org/10.1016/S0169-2070\(03\)00014-1](https://doi.org/10.1016/S0169-2070(03)00014-1)
- Kang Y, Wang D, Yang K (2021) A new INAR(1) process with bounded support for counts showing equidispersion. *Stat Pap* 62(2):745–767. <https://doi.org/10.1007/s00362-019-01111-0>
- Karlin S, H.E. T. (1975) *A first course in stochastic processes* (2nd). Academic, New York
- Klimko LA, Nelson PI (1978) On conditional least squares estimation for stochastic processes. *Ann Stat* 6(3):629–642. <https://doi.org/10.1214/aos/1176344207>
- McKenzie E (1985) Some simple models for discrete variate time series. *J Am Water Resour Assoc* 21(4):645–650. <https://doi.org/10.1111/j.1752-1688.1985.tb05379.x>
- Möller T, Silva M, Weiß C et al (2016) Self-exciting threshold binomial autoregressive processes. *Adv Stat Anal* 100(4):369–400. <https://doi.org/10.1007/s10182-015-0264-6>
- Nik S, Weiß CH (2021) Smooth-transition autoregressive models for time series of bounded counts. *Stoch Model* 37(4):568–588. <https://doi.org/10.1080/15326349.2021.1945934>
- Pedeli X, Davison AC, Fokianos K (2015) Likelihood estimation for the INAR(p) model by saddlepoint approximation. *J Am Stat Assoc* 110(511):1229–1238. <https://doi.org/10.1080/01621459.2014.983230>
- Ristić MM, Weiß CH, Janjić AD (2016) A binomial integer-valued ARCH model. *Int J Biostat* 12(2). <https://doi.org/10.1515/ijb-2015-0051>
- Scotto MG, Weiß CH, Silva ME et al (2014) Bivariate binomial autoregressive models. *J Multivar Anal* 125:233–251. <https://doi.org/10.1016/j.jmva.2013.12.014>
- Silva MED, Oliveira VL (2004) Difference equations for the higher-order moments and cumulants of the INAR(1) model. *J Time Ser Anal* 25(3):317–333. <https://doi.org/10.1111/j.1467-9892.2004.01685.x>

- Steutel F, van Harn K (1979) Discrete analogues of self-decomposability and stability. *Ann Probab* 7(5):893–899. <https://doi.org/10.1214/aop/1176994950>
- Wang C, Liu H, Yao JF et al (2014) Self-excited threshold poisson autoregression. *J Am Stat Assoc* 109(506):777–787. <https://doi.org/10.1080/01621459.2013.872994>
- Wang D, Cui S, Cheng J et al (2021) Statistical inference for the covariates-driven binomial AR(1) process. *Acta Math Appl Sin Engl Ser* 37:758–772. <https://doi.org/10.1007/s10255-021-1043-7>
- Wang X (2020) Variable selection for first-order Poisson integer-valued autoregressive model with covariables. *Aust N Z J Stat* 62:278–295. <https://doi.org/10.1111/anzs.12295>
- Weiß CH (2009) Monitoring correlated processes with binomial marginals. *J Appl Stat* 36(4):399–414. <https://doi.org/10.1080/02664760802468803>
- Weiß CH (2009) A new class of autoregressive models for time series of binomial counts. *Commun Stat Theory Methods* 38(4):447–460. <https://doi.org/10.1080/03610920802233937>
- Weiß CH, Pollett PK (2014) Binomial autoregressive processes with density dependent thinning. *J Time Ser Anal* 35(2):115–132. <https://doi.org/10.1002/jtsa.12054>
- Yang K, Wang D, Jia B et al (2018) An integer-valued threshold autoregressive process based on negative binomial thinning. *Stat Pap* 59(3):1131–1160. <https://doi.org/10.1007/S00362-016-0808-1>
- Yang K, Wang D, Li H (2018) Threshold autoregression analysis for finite range time series of counts with an application on measles data. *J Stat Comput Simul* 88(3):597–614. <https://doi.org/10.1080/00949655.2017.1400032>
- Yang K, Li H, Wang D et al (2021) Random coefficients integer-valued threshold autoregressive processes driven by logistic regression. *AStA Adv Stat Anal* 105:533–557. <https://doi.org/10.1007/s10182-020-00379-0>
- Yang K, Yu X, Zhang Q et al (2022) On MCMC sampling in self-exciting integer-valued threshold time series models. *Comput Stat Data Anal* 169(107):410. <https://doi.org/10.1016/j.csda.2021.107410>
- Yang K, Li A, Li H et al (2023) High-order self-excited threshold integer-valued autoregressive model: estimation and testing. *Commun Math Stat*. <https://doi.org/10.1007/s40304-022-00325-3>
- Zhang J, Wang D, Yang K et al (2020) A multinomial autoregressive model for finite-range time series of counts. *J Stat Plan Inference* 207:320–343. <https://doi.org/10.1016/j.jspi.2020.01.005>
- Zhang J, Wang J, Tai Z et al (2022) A study of binomial AR(1) process with an alternative generalized binomial thinning operator. *J Korean Stat Soc* 52:110–129. <https://doi.org/10.1007/s42952-022-00193-1>
- Zhang R, Wang D (2023) A new binomial autoregressive process with explanatory variables. *J Comput Appl Math* 420(114):814. <https://doi.org/10.1016/j.cam.2022.114814>
- Zhu R, Joe H (2006) Modelling count data time series with Markov processes based on binomial thinning. *J Time Ser Anal* 27(5):725–738. <https://doi.org/10.1111/j.1467-9892.2006.00485.x>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.