



# A new thinning-based INAR(1) process for underdispersed or overdispersed counts

Yao Kang<sup>1</sup> · Dehui Wang<sup>1</sup> · Kai Yang<sup>2</sup> · Yulin Zhang<sup>1</sup>

Received: 15 December 2018 / Accepted: 19 July 2019 / Published online: 1 January 2020  
© Korean Statistical Society 2020

## Abstract

Underdispersed and overdispersed phenomena are often observed in practice. To deal with these phenomena, we introduce a new thinning-based integer-valued autoregressive process. Some probabilistic and statistical properties of the process are obtained. The asymptotic normality of the estimators of the model parameters, using conditional least squares, weighted conditional least squares and modified quasi-likelihood methods, are presented. One overdispersed real-data example and one underdispersed real-data example are given to show the flexibility and superiority of the new model.

**Keywords** INAR(1) process · Overdispersion · Underdispersion · GSC thinning operator

**Mathematics Subject Classification** 62M10 · 62J20

## 1 Introduction

During the last three decades, an integer-valued autoregressive (of order 1) model, denoted by INAR(1), has been widely used in real-world applications, such as the insurance actuarial, reliability theory, medicine and social sciences. There is a huge literature on the INAR(1) model [see Weiß (2008b) and Scotto et al. (2015)]. For more details on the INAR(1) model, we refer to Jazi et al. (2012), Schweer and Weiß (2014), Bourguignon and Vasconcellos (2015), Barreto-Souza (2015, 2017), Li et al. (2015), Borges et al. (2016), Bourguignon and Weiß (2017), Kim and Lee (2017) and Bourguignon et al. (2019).

---

✉ Dehui Wang  
wangdh@jlu.edu.cn

<sup>1</sup> School of Mathematics, Jilin University, Changchun 130012, China

<sup>2</sup> School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China

The INAR(1) processes have been constructed by using thinning operators. The binomial thinning operator was originally proposed by Steutel and Van Harn (1979) and defined as

$$\alpha \circ X = \sum_{i=1}^X B_i, \quad (1)$$

where  $\{B_i\}$  is a sequence of iid Bernoulli( $\alpha$ ) random variables independent of  $X$ . Based on the binomial thinning operator, the Poisson INAR(1) (POINAR(1)) model was introduced by Alzaid and Al-Osh (1987) and defined by

$$X_t = \alpha \circ X_{t-1} + \epsilon_t, \quad t \geq 1,$$

where  $\{\epsilon_t\}$  is a sequence of iid Poisson random variables with mean  $\lambda(1 - \alpha)$ , uncorrelated with the past value of  $\{X_t\}$ . Moreover,  $X_0$  is assumed to follow the Poisson distribution with mean  $\lambda$ .

The POINAR(1) model has been widely used in practice due to its simplicity. For instance, Cossette et al. (2011) generalized the classical discrete time risk models which used the POINAR(1) model to describe the dependence (in time) between the claim frequencies. Freeland and McCabe (2004) applied the POINAR(1) model that exhibited the short-range dependence to analyse a data set.

While the POINAR(1) model is widely used, it has two main limitations in practice. First, the binomial thinning operator in the POINAR(1) model is not appropriate when the observed unit can generate more counting objects or produce more new random events. Second, the Poisson distribution suffers from the equidispersion requirement which can not explain underdispersion and overdispersion. To handle the first limitation, Ristić et al. (2009) introduced the negative binomial thinning operator

$$\alpha * X = \sum_{i=1}^X W_i, \quad (2)$$

where  $\{W_i\}$  is a sequence of iid Geometric( $\alpha/(1 + \alpha)$ ) random variables independent of  $X$ . The negative binomial thinning operator contains geometric counting series which can explain overdispersion. To deal with the second limitation, Ristić et al. (2009) introduced a negative binomial thinning INAR(1) model with the geometric marginal, i.e., NGINAR(1) process. Since the NGINAR(1) model captures overdispersion, during the past ten years, the NGINAR(1) model has become popular in some fields such as the reliability theory, medicine and reservoirs theory.

However, the NGINAR(1) model still has some drawbacks: (i) the counting series in the negative binomial thinning operator can not exhibit equidispersion and underdispersion, (ii) the geometric marginal distributions are not suitable for explaining underdispersion. The aim of this paper is to deal with the above two problems. For the first drawback, we propose a new thinning operator by using a more general discrete distribution. Our thinning operator has three attractive characteristics: (i) the counting

series in our thinning operator can exhibit equidispersion, overdispersion, underdispersion, zero inflation and zero deflation; (ii) the probability mass function (pmf) of the counting random variable in our thinning operator is a decreasing function, which is very useful to explain some practical problems; (iii) the counting random variable in our thinning operator can describe both the short and long tailed count data. To the best of our knowledge, there is no thinning operator that can capture all the above three features. For the second drawback, we propose a new thinning-based INAR(1) process to explain both overdispersion and underdispersion.

The contents of this paper are organized as follows. In Sect. 2, we construct a new INAR(1) process based on an alternative thinning operator. In Sect. 3, some probabilistic and statistical properties of the process are derived. In Sect. 4, the estimators of the model parameters are derived by using conditional least squares (CLS), weighted conditional least squares (WCLS) and modified quasi-likelihood (MQL) methods. Also, the asymptotic properties of the estimators are investigated. In Sect. 5, we present some simulation studies to investigate the performances of the proposed estimators. In Sect. 6, we apply the model to two real data sets. Section 7 contains a discussion on the higher-order model. All proofs are given in Appendix.

## 2 Construction of the new process

Gómez-Déniz et al. (2011) introduced a new discrete distribution taking non-negative integers  $\{0, 1, \dots\}$ . For convenience, we call it the Gómez–Déniza–Sarabia–Calderín–Ojeda (GSC) distribution. The pmf of the GSC distribution is given by

$$P_n = \Pr(N = n) = \frac{\log(1 - \alpha\theta^n) - \log(1 - \alpha\theta^{n+1})}{\log(1 - \alpha)}, \quad (3)$$

where  $\alpha < 1$ ,  $\alpha \neq 0$  and  $0 < \theta < 1$ . The moments and the moment generating function (mgf) of the GSC distribution are given as follows:

$$\begin{aligned} E(N) &= \frac{1}{\log(1 - \alpha)} \sum_{s=1}^{\infty} \log(1 - \alpha\theta^s), \\ E(N^2) &= \frac{1}{\log(1 - \alpha)} \sum_{s=1}^{\infty} (2s - 1) \log(1 - \alpha\theta^s), \\ E(t^N) &= \frac{1}{\log(1 - \alpha)} \sum_{s=0}^{\infty} t^s \log\left(\frac{1 - \alpha\theta^s}{1 - \alpha\theta^{s+1}}\right), \quad t \in [-1, 1]. \end{aligned}$$

There are four main advantages of the GSC distribution. Firstly, Gómez-Déniz et al. (2011) found that overdispersion and underdispersion are encountered depending on the values of the distribution parameters. In our study, we found that some parameter combinations can also lead to equidispersion, which means that the GSC distribution can be viewed as an alternative one to the Poisson distribution. To investigate the dispersion characteristic of  $GSC(\alpha, \theta)$  with varying values of  $\alpha$  and  $\theta$ , the index of

**Table 1** The indices of dispersion of  $GSC(\alpha, \theta)$  for different values of  $\alpha$  and  $\theta$

$\alpha$	-50	-7.855	-5.57	-1	0	0.3	0.7
$\theta$							
0.1	0.8895	0.9786	1	1.0777	1.1111	1.1247	1.1489
0.125	0.9279	1	1.0207	1.1038	1.1429	1.1592	1.1891
0.2	1.0626	1.0958	1.1124	1.1986	1.2500	1.2736	1.3198
0.25	1.1677	1.1798	1.1935	1.2765	1.3333	1.3610	1.4177

**Table 2** The zero inflation indices of  $GSC(\alpha, \theta)$  for different values of  $\alpha$  and  $\theta$

$\alpha$	-10	-5	0	0.3	0.7
$\theta$					
0.1	-0.0077	0	0.0058	0.0055	0.0046
0.25	0.0701	0.0599	0.0467	0.0435	0.0356
0.5	0.9593	0.7148	0.3591	0.3218	0.2517

dispersion of the GSC distribution;  $Var(N)/E(N)$ , is shown in Table 1. Secondly, the GSC distribution can be used to explain varying degrees of the zero-inflated and slightly zero-deflated phenomena. To illustrate it, we introduce a zero inflation index;  $z_i = p_0 \exp(\mu) - 1$  (see Weiß et al. 2019), where  $p_0$  is the proportion of 0's and  $\mu$  is the mean.  $z_i > 0$  means that the distribution suffers from zero inflation and  $z_i < 0$  means the random variable is zero-deflated. The zero inflation indices of  $GSC(\alpha, \theta)$  with varying values of  $\alpha$  and  $\theta$  are shown in Table 2. Thirdly, the GSC distribution represents a general family of distributions. As pointed out by Gómez-Déniz et al. (2011), the GSC distribution can be viewed as a compound Poisson distribution, a specific mixed geometric distribution and a specific mixed Poisson distribution. Furthermore, the GSC distribution can be viewed as a possible alternative one to the negative binomial, generalized Poisson, hyper-Poisson, Poisson-inverse Gaussian distribution, different generalizations of the geometric distribution that have been discussed in the statistical literature. Finally, Gómez-Déniz et al. (2011) pointed out that the pmf (3) is a decreasing function and has the ability to describe the short and long tailed counts.

Although the GSC distribution has some interesting properties, it still has a significant disadvantage: as pointed out by a referee, the GSC distribution does not have explicit formulae for the mean and variance, which leads to some difficulties when it is applied to the statistical model. To analyse this drawback, we discuss it from two aspects. Firstly, we prove that the infinite sum in the mean and variance expressions are convergent (see the top of Appendix). Secondly, the infinite sum in the mean and variance of the GSC distribution can be approximated very precisely by the corresponding finite sum. This conclusion can be supported by Table 3, i.e., the mean and variance of  $GSC(\alpha, \theta)$  for different values of  $\alpha$  and  $\theta$  with varying upper limit in the finite sum. From Table 3, we find that the approximation is precise enough for each parameter combination when the finite sum as the indices running  $s = 1, \dots, 200$ . Based on these discussions, we conclude that the inexistence of explicit formulae for

**Table 3** The means (above) and variances (below) of  $GSC(\alpha, \theta)$  for different values of  $\alpha$  and  $\theta$  with varying upper limit in the infinite series

$s$	10	20	50	100	200	500	1000
$(\alpha, \theta)$							
(-5, 0.1)	0.2566	0.2566	0.2566	0.2566	0.2566	0.2566	0.2566
	0.2583	0.2583	0.2583	0.2583	0.2583	0.2583	0.2583
(-5, 0.9)	7.5016	11.3680	13.9388	14.0668	14.0674	14.0674	14.0674
	11.6160	49.8558	131.8660	143.4529	143.5802	143.5802	143.5802
(0.3, 0.1)	0.0948	0.0948	0.0948	0.0948	0.0948	0.0948	0.0948
	0.1066	0.1066	0.1066	0.1066	0.1066	0.1066	0.1066
(0.3, 0.9)	5.4810	7.2605	8.1499	8.1887	8.1889	8.1889	8.1889
	14.4548	42.0325	78.3814	82.3472	82.3879	82.3879	82.3879
(0.7, 0.1)	0.0668	0.0668	0.0668	0.0668	0.0668	0.0668	0.0668
	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767
(0.7, 0.9)	4.5857	5.8787	6.5012	6.5280	6.5281	6.5281	6.5281
	14.4119	37.2870	64.5604	67.3919	67.4205	67.4205	67.4205

the mean and variance can be overcome by approximating the infinite sum by the finite sum.

To get a better fit for a serially dependent count data, one should identify the dispersion behavior of the counts before choosing a suitable INAR(1) process. The most popular methods to identify the dispersion behavior of the counts is a test proposed by Schweer and Weiß (2014). Consider the null hypothesis  $\mathcal{H}_0: X_1, \dots, X_n$  stem from an equidispersed POINAR(1) process ( $I_d = 1$ ) against the alternative of an overdispersed (or underdispersed) marginal distribution. Let  $z_{1-\beta}$  be the quantile of the  $(1 - \beta)$ -quantile of the  $N(0, 1)$ -distribution. We reject the null hypothesis  $\mathcal{H}_0$ : equidispersion on significance level  $\beta$  in favor of alternative hypothesis  $\mathcal{H}_1$ : overdispersion (or underdispersion) if

$$\hat{I}_d > 1 + z_{1-\beta} \sqrt{\frac{2(1 + \hat{\rho}_X^2(1))}{n(1 - \hat{\rho}_X^2(1))}} \left( \text{or } \hat{I}_d < 1 + z_{\beta} \sqrt{\frac{2(1 + \hat{\rho}_X^2(1))}{n(1 - \hat{\rho}_X^2(1))}} \right),$$

where  $\hat{I}_d = \sum_{t=1}^n (X_t - \bar{X})^2 / \sum_{t=1}^n X_t$ ,  $\bar{X} = (1/n) \sum_{t=1}^n X_t$  and  $\hat{\rho}_X(1)$  is the first-order autocorrelation coefficient of  $X_1, \dots, X_n$ .

After identifying the dispersion behavior of the counts, one may choose a suitable existing INAR(1) model to fit the data. On one hand, the binomial thinning INAR(1) models with different innovation structures are natural choices. For example, Bourguignon and Vasconcellos (2015), Kim and Lee (2017), Bourguignon et al. (2019) introduced the binomial thinning INAR(1) processes with the power series, Katz family, double Poisson and generalized Poisson innovations, respectively. The above models are able to handle equidispersion, underdispersion and overdispersion. On the other hand, the negative binomial thinning INAR(1) models with different marginal distributions are also commonly used. As we mentioned before, the NGINAR(1) model

is very popular when the overdispersed counts are suffered. Barreto-Souza (2015) proposed a negative binomial thinning INAR(1) process with the zero-modified geometric marginal to account for underdispersion and overdispersion. However, new thinning-based INAR(1) model is still needed. This statement can be explained from two aspects. Firstly, the binomial and negative binomial thinning operators both have some limitations: (i) the binomial thinning operator is not suitable when the observed unit can generate more counting objects or produce more new random events; (ii) the negative binomial thinning operator can not exhibit equidispersion and underdispersion. Secondly, due to the complexity and diversity of the practical application, the counting series in the thinning operator are expected to have the ability to explain as many data characteristics as possible.

Based on the need for the new thinning operator and the attractive advantages of the GSC distribution, we use this distribution to create a GSC thinning operator which is defined by

$$\alpha \diamond X := \sum_{j=1}^X W_j, \tag{4}$$

where  $\{W_j\}$  is a sequence of iid  $GSC(\alpha, \exp\{-|\alpha|\})$  random variables,  $E(W_j) = \phi$ ,  $Var(W_j) = \beta$ ,  $\{W_j\}$  and  $X$  are independent. The proposed thinning operator (4) not only can overcome the shortcomings of the binomial and negative binomial thinning operators (1) and (2), but also has the ability to describe many data characteristics. To be specific, the GSC thinning operator can capture the feature that the observed unit may generate more counting objects or produce more new random events and the counting series in our thinning operator can show equidispersion, overdispersion and underdispersion. Besides, the counting series in our thinning operator can describe the zero-inflated, zero-deflated, short tailed and long tailed characteristics.

We now introduce GSC thinning-based INAR(1) process, as follows:

**Definition 1** An INAR(1) model based on the GSC thinning operator, denoted by GSCINAR(1), is defined by the following difference equation:

$$X_t = \alpha \diamond X_{t-1} + \epsilon_t, \quad t \geq 1, \tag{5}$$

where  $\{W_j\}$  is a sequence of iid  $GSC(\alpha, \exp\{-|\alpha|\})$  random variables with the finite mean  $\phi$  and variance  $\beta$ ,  $\alpha < 1$ ,  $\alpha \neq 0$ . Here, we write  $\phi = \frac{1}{\log(1-\alpha)} \sum_{s=1}^{\infty} \log(1 - \alpha \exp\{-s|\alpha|\})$  and  $\beta = \frac{1}{\log(1-\alpha)} \sum_{s=1}^{\infty} (2s - 1)\log(1 - \alpha \exp\{-s|\alpha|\}) - \phi^2$ .  $\{\epsilon_t\}$  is an innovation sequence of iid non-negative integer-valued random variables, uncorrelated with the past values of  $\{X_t\}$ . Let  $\mu_\epsilon = E(\epsilon_t)$ ,  $\sigma_\epsilon^2 = Var(\epsilon_t)$  (we assume that they exist).

### 3 Properties of GSCINAR(1) process

In this section, we consider some properties of GSCINAR(1) process.

**Proposition 1** Suppose  $\{X_t\}$  is a stationary process satisfying (5). Then for  $t \geq 1$ ,

- (i)  $E(X_t|X_{t-1}) = \phi X_{t-1} + \mu_\epsilon$ ,
- (ii)  $E(X_t) = \mu_\epsilon / (1 - \phi)$ ,
- (iii)  $\text{Var}(X_t|X_{t-1}) = \beta X_{t-1} + \sigma_\epsilon^2$ ,
- (iv)  $\text{Var}(X_t) = [\beta \mu_\epsilon + \sigma_\epsilon^2(1 - \phi)] / [(1 - \phi)^2(1 + \phi)]$ ,
- (v)  $\rho_k = \text{Corr}(X_{t+k}, X_t) = \phi^k, k = 1, 2, \dots$ ,

where  $\phi$  and  $\beta$  are given in Definition 1.

**Remark 1** (i) Proposition 1(i) shows that the GSCINAR(1) model is a member of the non-Gaussian conditional linear AR(1) models discussed by Grunwald et al. (2000).

(ii) The index of dispersion of  $\{X_t\}$  is given by

$$I_X := \frac{\text{Var}(X_t)}{E(X_t)} = \frac{\beta \mu_\epsilon + \sigma_\epsilon^2(1 - \phi)}{\mu_\epsilon(1 + \phi)(1 - \phi)}.$$

Following Li et al. (2015), the existence of the strict stationary and ergodic GSCINAR(1) process can be established in the following theorem.

**Theorem 1** If  $0 < \phi < 1$ , then there exists a unique strictly stationary integer-valued random series  $\{X_t\}$  satisfying

$$X_t = \alpha \diamond X_{t-1} + \epsilon_t, \quad t \geq 1,$$

$\text{Cov}(X_s, \epsilon_t) = 0$  for  $s < t$ . Furthermore, the process is an ergodic process.

## 4 Estimation of the unknown parameters

Suppose  $\{X_t\}$  is a strictly stationary and ergodic solution of model (5). Our task is to estimate the parameter  $\eta = (\alpha, \mu_\epsilon)$  from a sample  $(X_1, X_2, \dots, X_n)$ . Three different methods of parameter estimation, the CLS, WCLS and MQL, are applied. The reason why we take these approaches is that they do not require specifying the exact family of distributions for the innovations.

### 4.1 Conditional least squares estimator

The CLS estimator  $\hat{\eta}_{CLS} = (\hat{\alpha}_{CLS}, \hat{\mu}_{\epsilon_{CLS}})'$  of  $\eta$  is obtained by minimizing the expression

$$\begin{aligned} Q_1(\eta) &:= \sum_{t=1}^n (X_t - \phi X_{t-1} - \mu_\epsilon)^2 \\ &= \sum_{t=1}^n \left( X_t - \frac{X_{t-1}}{\log(1 - \alpha)} \sum_{s=1}^{\infty} \log(1 - \alpha \exp\{-s|\alpha|\}) - \mu_\epsilon \right)^2. \end{aligned} \quad (6)$$

The following result establishes the asymptotic distribution of  $\widehat{\eta}_{CLS}$ . For convenience, write

$$V_{\omega} = \begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix}, \quad H_{\omega} = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix},$$

where

$$\begin{aligned} V_{11} &= E\{[\omega(X_0)(\dot{\phi}X_0)(X_1 - \phi X_0 - \mu_{\epsilon})]^2\}, \\ V_{22} &= E\{[\omega(X_0)(X_1 - \phi X_0 - \mu_{\epsilon})]^2\}, \\ V_{12} &= E[\omega^2(X_0)(\dot{\phi}X_0)(X_1 - \phi X_0 - \mu_{\epsilon})^2] \\ H_{11} &= E[\omega(X_0)(\dot{\phi}X_0)^2], \quad H_{12} = E[\omega(X_0)(\dot{\phi}X_0)], \\ H_{22} &= E[\omega(X_0)], \quad \dot{\phi} = \frac{\partial \phi}{\partial \alpha} = \frac{\sum_{s=1}^{\infty} \log(1 - \alpha \exp\{-s|\alpha|\})}{(1 - \alpha) \log^2(1 - \alpha)} \\ &\quad + \sum_{s=1}^{\infty} \frac{(s|\alpha| - 1) \exp\{-s|\alpha|\}}{\log(1 - \alpha)(1 - \alpha \exp\{-s|\alpha|\})}, \end{aligned}$$

$\omega(\cdot)$  is a weight function. It can be verified that  $H_{\omega}$  is a invertible matrix.

**Theorem 2** *Suppose  $E|X_t|^4 < \infty$ . Then, we have*

$$\sqrt{n}(\widehat{\eta}_{CLS} - \eta) \xrightarrow{d} N(\mathbf{0}, H_{CLS}^{-1} V_{CLS} H_{CLS}^{-1}),$$

where  $V_{CLS}$  and  $H_{CLS}$  are given by  $V_{\omega}$  and  $H_{\omega}$ , with  $\omega(X_0) = 1$ .

### 4.2 Weighted conditional least squares estimator

In general, the CLS estimator is not asymptotically efficient. To improve the efficiency, we consider the WCLS estimator as an alternative one to the CLS estimator. In this section, we focus on the WCLS method with a known weight function. The WCLS estimator  $\widehat{\eta}_{WCLS} = (\widehat{\alpha}_{WCLS}, \widehat{\mu}_{\epsilon_{WCLS}})'$  can be obtained by minimizing

$$\begin{aligned} Q_2(\eta) &:= \sum_{t=1}^n \omega(X_{t-1})(X_t - \phi X_{t-1} - \mu_{\epsilon})^2 \\ &= \sum_{t=1}^n \omega(X_{t-1}) \left( X_t - \frac{X_{t-1}}{\log(1 - \alpha)} \sum_{s=1}^{\infty} \log(1 - \alpha \exp\{-s|\alpha|\}) - \mu_{\epsilon} \right)^2, \quad (7) \end{aligned}$$

where  $\omega(X_{t-1})$  is a suitably chosen weight function. A natural choice of  $\omega(X_{t-1})$  may be

$$\omega(X_{t-1}) = \frac{1}{X_{t-1} + c_1}, \quad (8)$$

where  $c_1$  is a positive constant.



The following result establishes the asymptotic distribution of  $\widehat{\eta}_{WCLS}$ . The proof is similar to the proof of Theorem 2 and we omit it.

**Theorem 3** *Suppose  $E|X_t|^4 < \infty$ . Then, we have*

$$\sqrt{n}(\widehat{\eta}_{WCLS} - \eta) \xrightarrow{d} N(\mathbf{0}, \mathbf{H}_{WCLS}^{-1} \mathbf{V}_{WCLS} \mathbf{H}_{WCLS}^{-1}),$$

where  $\mathbf{V}_{WCLS}$  and  $\mathbf{H}_{WCLS}$  are given by  $\mathbf{V}_\omega$  and  $\mathbf{H}_\omega$ , with  $\omega(X_0) = 1/(X_0 + c_1)$ .

### 4.3 Modified quasi-likelihood estimator

Let  $\tau = (\alpha, \sigma_\epsilon^2)'$ . Recall that, from Proposition 1(iii), the expression for the one-step conditional variance is

$$V_\tau(X_t|X_{t-1}) = \text{Var}(X_t|X_{t-1}) = \beta X_{t-1} + \sigma_\epsilon^2,$$

where  $\beta$  is given in Definition 1. The MQL estimator  $\widehat{\eta}_{MQL} = (\widehat{\alpha}_{MQL}, \widehat{\mu}_{\epsilon_{MQL}})'$  can be obtained by minimizing

$$\begin{aligned} Q_3(\eta) &:= \sum_{t=1}^n V_{\widehat{\tau}}^{-1}(X_t|X_{t-1})(X_t - \phi X_{t-1} - \mu_\epsilon)^2 \\ &= \sum_{t=1}^n V_{\widehat{\tau}}^{-1}(X_t|X_{t-1}) \left( X_t - \frac{X_{t-1}}{\log(1 - \alpha)} \sum_{s=1}^\infty \log(1 - \alpha \exp\{-s|\alpha|\}) - \mu_\epsilon \right)^2, \end{aligned} \tag{9}$$

where  $\widehat{\tau}$  is a consistent estimator of  $\tau$ . Note that  $\widehat{\eta}_{CLS}$  is a consistent estimator of  $\eta$  (see Theorem 2), while the consistent estimator of  $\sigma_\epsilon^2$  can be obtained, as follows:

The first method is based on the moment estimator

$$\widehat{\sigma}_\epsilon^2 = (1 - \widehat{\phi}^2)s^2 - \widehat{\beta}\overline{X}, \tag{10}$$

where  $s^2 = \sum_{t=1}^n (X_t - \overline{X})^2 / (n - 1)$ ,  $\overline{X} = \sum_{t=1}^n X_t / n$ ,  $\widehat{\phi} = \frac{1}{\log(1 - \widehat{\alpha})} \sum_{s=1}^\infty \log(1 - \widehat{\alpha} \exp\{-s|\widehat{\alpha}|\})$ ,  $\widehat{\beta} = \frac{1}{\log(1 - \widehat{\alpha})} \sum_{s=1}^\infty (2s - 1) \log(1 - \widehat{\alpha} \exp\{-s|\widehat{\alpha}|\}) - \widehat{\phi}^2$ ,  $\widehat{\alpha}$  is the CLS estimator of  $\alpha$ .

The second method is based on the two-step CLS method which has been discussed by Karlsen and Tjøstheim (1988). Let

$$S(\alpha, \mu_\epsilon, \sigma_\epsilon^2) = \sum_{t=1}^n [\{X_t - E(X_t|X_{t-1})\}^2 - \text{Var}(X_t|X_{t-1})]^2.$$

Then, the consistent estimator of  $\sigma_\epsilon^2$  may be obtained by minimizing  $S(\widehat{\alpha}_{CLS}, \widehat{\mu}_{\epsilon_{CLS}}, \sigma_\epsilon^2)$  with respect to  $\sigma_\epsilon^2$ .

Following Zheng et al. (2007), we establish the asymptotic distribution of  $\widehat{\eta}_{MQL}$  in the following theorem.

**Theorem 4** Suppose  $E|X_t|^4 < \infty$ . Then, we have

$$\sqrt{n}(\widehat{\eta}_{MQL} - \eta) \xrightarrow{d} N(\mathbf{0}, \mathbf{H}_{MQL}^{-1} \mathbf{V}_{MQL} \mathbf{H}_{MQL}^{-1}),$$

where  $\mathbf{V}_{MQL}$  and  $\mathbf{H}_{MQL}$  are given by  $\mathbf{V}_\omega$  and  $\mathbf{H}_\omega$ , with  $\omega(X_0) = \mathbf{V}_\tau^{-1}(X_1|X_0)$ .

### 5 Simulation studies

In this section, the estimators described earlier are compared by simulations. Consider

$$X_t = \alpha \diamond X_{t-1} + \epsilon_t, \quad t \geq 1,$$

where  $\{\epsilon_t\}$  is a sequence of iid Poisson random variables (Model A) with mean  $\mu_\epsilon$  or iid generalized Poisson (GP) random variables (Model B) with  $\mu_\epsilon = \lambda_\epsilon / (1 - \kappa_\epsilon)$ .

**Remark 2** A random variable  $X$  has a GP distribution with parameters  $\lambda$  and  $\kappa$ , which we denote by  $GP(\lambda, \kappa)$ , if its pmf is

$$P(X = x) = \begin{cases} \lambda(\lambda + \kappa x)^{x-1} e^{-(\lambda + \kappa x)} / x!, & x = 0, 1, 2, \dots, \\ 0, & \text{for } x > m \text{ if } \kappa < 0, \end{cases}$$

where  $\lambda > 0$ ,  $\max(-1, -\lambda/m) < \kappa < 1$ , and  $m (\leq 4)$  is the largest positive integer for which  $\lambda + \kappa m > 0$  when  $\kappa < 0$ . The mean and variance of  $GP(\lambda, \kappa)$  are

$$E(X) = \frac{\lambda}{1 - \kappa}, \quad \text{Var}(X) = \frac{\lambda}{(1 - \kappa)^3}.$$

In the simulation, we generate the GSCINAR(1) sample with the sample size equals to  $n + 1000$  and discard the first 1000 observations. We generate the data from the models and set the sample sizes  $n = 300, 500, 800, 1000$ . The true values of the parameters are:

Model A: (A1)  $(\alpha, \mu_\epsilon) = (-2, 2)$ ; (A2)  $(\alpha, \mu_\epsilon) = (-1.5, 2.5)$ ;

Model B: (B1)  $(\alpha, \lambda_\epsilon, \kappa_\epsilon) = (0.8, 0.5, 0.5)$ ; (B2)  $(\alpha, \lambda_\epsilon, \kappa_\epsilon) = (0.85, 0.75, 0.5)$ .

Figure 1 is the sample paths from Models A and B. Table 4 lists some statistics of Models A and B including the mean, variance, first-order autocorrelation coefficient (ACF(1)) and zero probability ( $p_0$ ). Specially, the zero probability is computed from the average percentages of zeros in time series of length 5000 generated from the corresponding models. The average is obtained based on 1000 replications.

To compare the three methods, we calculate the mean squared error (MSE) and standard deviation (SD) based on  $m = 1000$  replications for each combinations;  $MSE = \frac{1}{m} \sum_{k=1}^m (\widehat{\alpha}_k - \alpha)^2$ ,  $SD = \sqrt{\frac{1}{m-1} \sum_{k=1}^m (\widehat{\alpha}_k - \bar{\alpha})^2}$ , where  $\widehat{\alpha}_k$  is the estimator of  $\alpha$  at the  $k$ th replication and  $\bar{\alpha} = \frac{1}{m} \sum_{k=1}^m \widehat{\alpha}_k$ . For simplification of the computation,

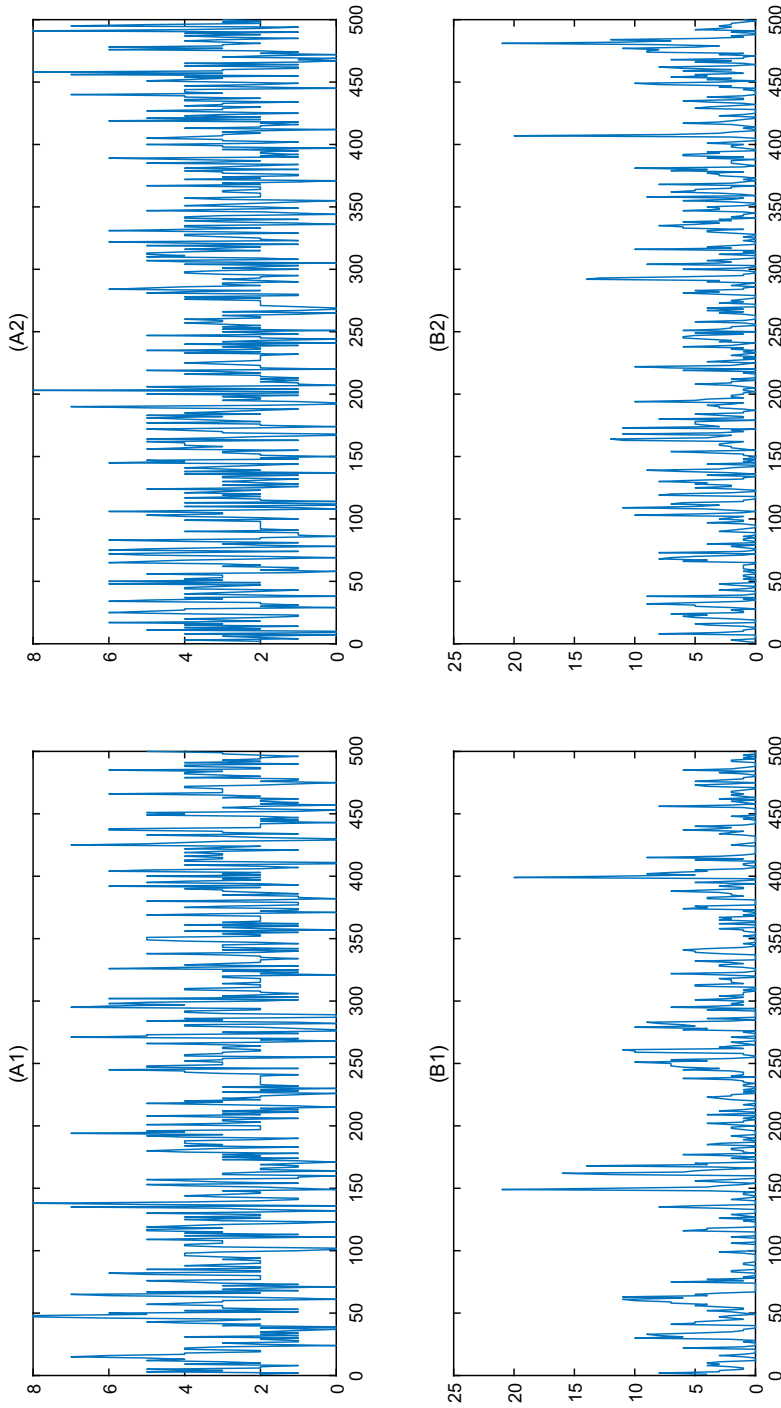


Fig. 1 The sample paths of Models (A1)–(B2) with the sample size  $n = 500$

**Table 4** Some statistics for Models (A1)–(B2)

Model	Mean	Variance	ACF(1)	Zero probability
(A1)	2.6882	2.9412	0.2560	0.0759
(A2)	4.2877	5.6559	0.4169	0.0226
(B1)	1.8877	7.4112	0.4703	0.4134
(B2)	2.4570	9.2682	0.3895	0.3002

**Table 5** The MSEs and SDs of the WCLS estimators with different weights and the sample size  $n = 1000$

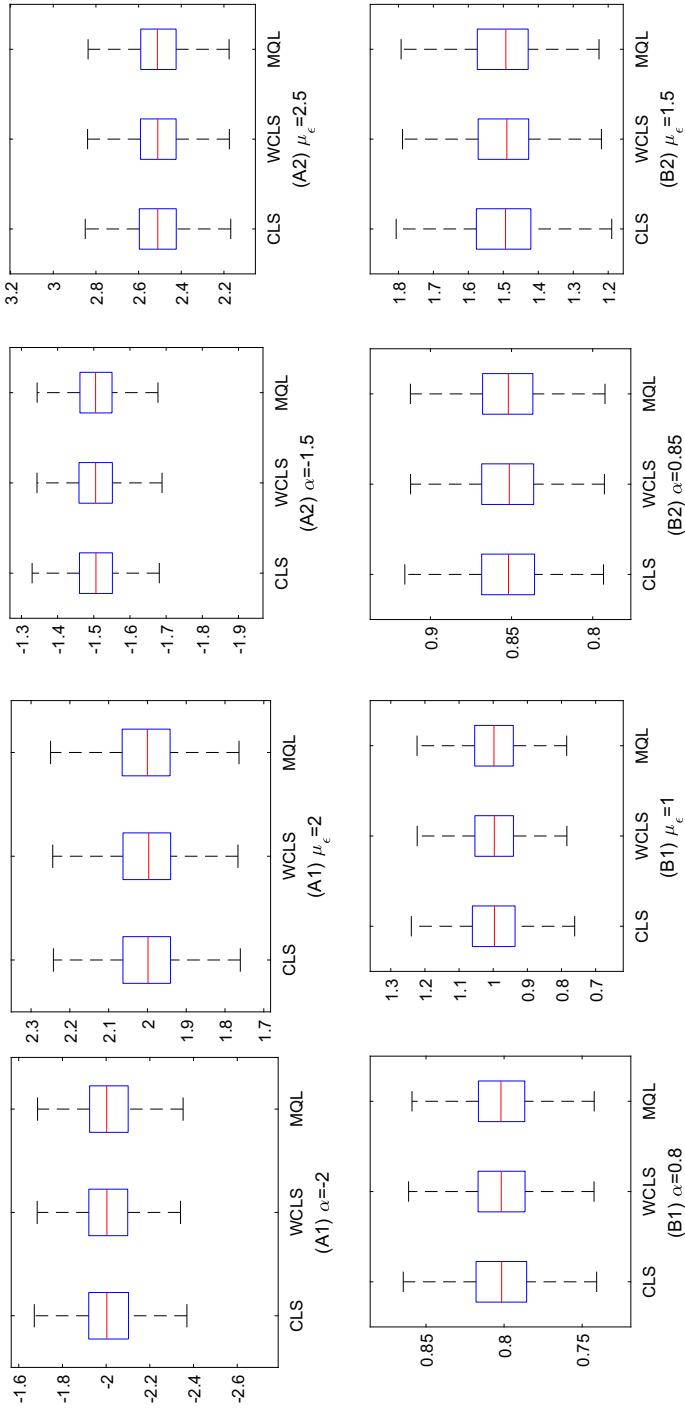
Model	Parameter	$c_1 = 1$		$c_1 = 3$		$c_1 = 5$		$c_1 = 7$		$c_1 = 9$	
		MSE	SD	MSE	SD	MSE	SD	MSE	SD	MSE	SD
(A1)	$\alpha = -2$	0.0241	0.1540	0.0197	0.1394	0.0193	0.1379	0.0192	0.1378	0.0193	0.1379
	$\mu_\epsilon = 2$	0.0104	0.1019	0.0091	0.0954	0.0090	0.0948	0.0090	0.0948	0.0090	0.0949
(A2)	$\alpha = -1.5$	0.0057	0.0754	0.0050	0.0707	0.0050	0.0707	0.0051	0.0710	0.0051	0.0713
	$\mu_\epsilon = 2.5$	0.0192	0.1386	0.0170	0.1305	0.0169	0.1302	0.0170	0.1306	0.0172	0.1312
(B1)	$\alpha = 0.8$	0.0005	0.0222	0.0005	0.0212	0.0005	0.0212	0.0005	0.0213	0.0005	0.0214
	$\mu_\epsilon = 1$	0.0072	0.0849	0.0068	0.0824	0.0068	0.0824	0.0068	0.0827	0.0069	0.0830
(B2)	$\alpha = 0.85$	0.0006	0.0247	0.0005	0.0224	0.0005	0.0219	0.0005	0.0217	0.0005	0.0216
	$\mu_\epsilon = 1.5$	0.0128	0.1133	0.0115	0.1074	0.0113	0.1065	0.0113	0.1063	0.0113	0.1064

all the infinite sum in (6), (7) and (9) are approximated by the corresponding finite sum running  $s = 1, \dots, 200$ .

To choose a more suitable weight for the WCLS method, we firstly compare the WCLS methods with different weights. We suppose that  $c_1$  in (8) equals to 1, 3, 5, 7 and 9, respectively. The simulation results are summarized in Table 5, which indicates that there is no significant difference among the five weights and a little better estimator can be obtained when  $c_1 = 3, 5, 7, 9$ . In the following, we set  $c_1 = 3$ .

From Table 6, the MSEs and SDs of the estimators decrease as the sample size  $n$  increases, as expected. This finding can be supported by the box plots shown in Fig. 2 (the box plots are symmetric and centered around the true parameter value). Next, we compare the three methods by observing the MSEs and SDs in Table 6. We find that the WCLS and MQL methods perform better than the CLS method. The smaller SDs indicate that the WCLS method improves the efficiency of the estimation and the weight function is satisfactory. The WCLS and MQL methods give the similar results in most cases and the MQL method can be a little better than the WCLS method. Figure 3 shows the QQ plots of the CLS, WCLS and MQL estimators for Model (B2) with the sample size  $n = 1000$ , which indicates that the CLS, WCLS and MQL estimators are asymptotically normal for all the parameters. Similar results can be obtained for all parameter combinations and the figures are also omitted.

To further compare the three methods, a contaminated model is considered here.



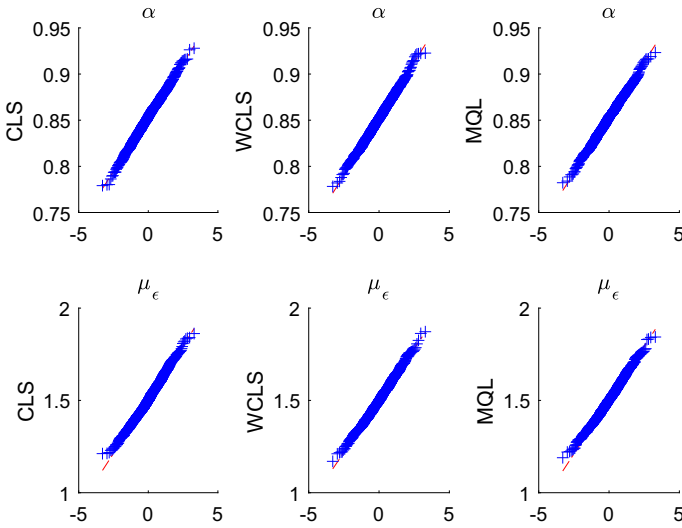
**Fig. 2** Box plots from 1000 simulated estimators of  $\alpha$  and  $\mu_\epsilon$  with the sample size  $n = 1000$ . (A1)  $\alpha = -2$ , (A1)  $\mu_\epsilon = 2$ , (A2)  $\alpha = -1.5$ , (A2)  $\mu_\epsilon = 2.5$ , (B1)  $\alpha = 0.8$ , (B1)  $\mu_\epsilon = 1$ , (B2)  $\alpha = 0.85$ , (B2)  $\mu_\epsilon = 1.5$

**Table 6** The MSEs and SDs of the estimators

Model	Sample size	Parameter	CLS		WCLS		MQL	
			MSE	SD	MSE	SD	MSE	SD
(A1)	$n = 300$	$\alpha = -2$	0.0797	0.2777	0.0794	0.2773	0.0765	0.2716
		$\mu_\epsilon = 2$	0.0281	0.1674	0.0281	0.1674	0.0272	0.1648
(A1)	$n = 500$	$\alpha = -2$	0.0415	0.2029	0.0399	0.1988	0.0394	0.1974
		$\mu_\epsilon = 2$	0.0176	0.1327	0.0169	0.1302	0.0167	0.1292
(A1)	$n = 800$	$\alpha = -2$	0.0257	0.1577	0.0252	0.1561	0.0245	0.1538
		$\mu_\epsilon = 2$	0.0109	0.1042	0.0107	0.1031	0.0105	0.1020
(A1)	$n = 1000$	$\alpha = -2$	0.0193	0.1388	0.0188	0.1369	0.0184	0.1355
		$\mu_\epsilon = 2$	0.0086	0.0926	0.0083	0.0913	0.0082	0.0906
(A2)	$n = 300$	$\alpha = -1.5$	0.0205	0.1397	0.0189	0.1345	0.0190	0.1341
		$\mu_\epsilon = 2.5$	0.0618	0.2466	0.0575	0.2380	0.0573	0.2370
(A2)	$n = 500$	$\alpha = -1.5$	0.0113	0.1049	0.0108	0.1025	0.0107	0.1016
		$\mu_\epsilon = 2.5$	0.0363	0.1897	0.0338	0.1830	0.0335	0.1820
(A2)	$n = 800$	$\alpha = -1.5$	0.0073	0.0847	0.0067	0.0808	0.0066	0.0805
		$\mu_\epsilon = 2.5$	0.0234	0.1528	0.0214	0.1461	0.0212	0.1453
(A2)	$n = 1000$	$\alpha = -1.5$	0.0047	0.0687	0.0042	0.0647	0.0042	0.0646
		$\mu_\epsilon = 2.5$	0.0175	0.1324	0.0158	0.1256	0.0157	0.1253
(B1)	$n = 300$	$\alpha = 0.8$	0.0022	0.0457	0.0018	0.0424	0.0020	0.0443
		$\mu_\epsilon = 1$	0.0248	0.1573	0.0230	0.1518	0.0232	0.1523
(B1)	$n = 500$	$\alpha = 0.8$	0.0011	0.0332	0.0009	0.0302	0.0009	0.0302
		$\mu_\epsilon = 1$	0.0160	0.1262	0.0136	0.1167	0.0137	0.1169
(B1)	$n = 800$	$\alpha = 0.8$	0.0007	0.0260	0.0006	0.0242	0.0006	0.0241
		$\mu_\epsilon = 1$	0.0097	0.0983	0.0087	0.0935	0.0087	0.0932
(B1)	$n = 1000$	$\alpha = 0.8$	0.0006	0.0237	0.0005	0.0213	0.0005	0.0213
		$\mu_\epsilon = 1$	0.0080	0.0895	0.0071	0.0845	0.0071	0.0845
(B2)	$n = 300$	$\alpha = 0.85$	0.0017	0.0407	0.0018	0.0425	0.0016	0.0397
		$\mu_\epsilon = 1.5$	0.0405	0.2007	0.0402	0.2002	0.0385	0.1956
(B2)	$n = 500$	$\alpha = 0.85$	0.0010	0.0320	0.0012	0.0340	0.0010	0.0311
		$\mu_\epsilon = 1.5$	0.0241	0.1552	0.0244	0.1563	0.0233	0.1524
(B2)	$n = 800$	$\alpha = 0.85$	0.0007	0.0257	0.0007	0.0258	0.0006	0.0248
		$\mu_\epsilon = 1.5$	0.0147	0.1213	0.0144	0.1199	0.0139	0.1177
(B2)	$n = 1000$	$\alpha = 0.85$	0.0006	0.0236	0.0005	0.0227	0.0005	0.0223
		$\mu_\epsilon = 1.5$	0.0129	0.1136	0.0122	0.1103	0.0120	0.1095

**Definition 2** (*Contaminated GSCINAR(1) Model*). A stochastic process  $(Y_k)_{k \in \mathbb{Z}_+}$  is called a contaminated GSCINAR(1) model if

$$Y_k = X_k + \xi_k c_2, \quad k \in \mathbb{Z}_+, \tag{11}$$



**Fig. 3** QQ plots of the CLS, WCLS and MQL estimators for Model (B2) with the sample size  $n = 1000$

**Table 7** The MSEs and SDs of the estimators in the contaminated situation with the sample size  $n = 1000$

Model	Contamination percentage	Parameter	CLS		WCLS		MQL	
			MSE	SD	MSE	SD	MSE	SD
(A1)	$\delta = 0.1$	$\alpha = -2$	0.0259	0.1519	0.0257	0.1515	0.0251	0.1490
		$\mu_\epsilon = 2$	0.0200	0.0976	0.0200	0.0976	0.0198	0.0960
(A1)	$\delta = 0.2$	$\alpha = -2$	0.0297	0.1532	0.0307	0.1535	0.0293	0.1502
		$\mu_\epsilon = 2$	0.0482	0.1008	0.0495	0.1005	0.0487	0.0988
(A2)	$\delta = 0.1$	$\alpha = -1.5$	0.0061	0.0751	0.0057	0.0716	0.0056	0.0711
		$\mu_\epsilon = 2.5$	0.0277	0.1394	0.0266	0.1332	0.0264	0.1323
(A2)	$\delta = 0.2$	$\alpha = -1.5$	0.0074	0.0770	0.0071	0.0744	0.0070	0.0737
		$\mu_\epsilon = 2.5$	0.0504	0.1387	0.0499	0.1333	0.0497	0.1321
(B1)	$\delta = 0.1$	$\alpha = 0.8$	0.0006	0.0239	0.0005	0.0217	0.0005	0.0214
		$\mu_\epsilon = 1$	0.0126	0.0903	0.0126	0.0835	0.0123	0.0823
(B1)	$\delta = 0.2$	$\alpha = 0.8$	0.0007	0.0248	0.0007	0.0223	0.0007	0.0219
		$\mu_\epsilon = 1$	0.0265	0.0974	0.0305	0.0904	0.0291	0.0883
(B2)	$\delta = 0.1$	$\alpha = 0.85$	0.0006	0.0246	0.0006	0.0236	0.0006	0.0229
		$\mu_\epsilon = 1.5$	0.0187	0.1171	0.0188	0.1127	0.0178	0.1107
(B2)	$\delta = 0.2$	$\alpha = 0.85$	0.0006	0.0245	0.0006	0.0230	0.0006	0.0224
		$\mu_\epsilon = 1.5$	0.0352	0.1184	0.0387	0.1162	0.0359	0.1127

where  $(X_k)_{k \in \mathbb{Z}_+}$  is a GSCINAR(1) process given by Definition 1.  $c_2$  is a positive integer and it represents the contamination’s size.  $(\delta_k)_{k \in \mathbb{Z}_+}$  is a sequence of iid Bernoulli( $\delta$ ) random variables. It is obvious to see that the contamination percentage is  $\delta$ .  $(X_k)_{k \in \mathbb{Z}_+}$  and  $(\xi_k)_{k \in \mathbb{Z}_+}$  are independent.

For comparison, we give the simulation results of Models A and B with the sample size  $n = 1000$  and different contamination percentages  $\delta = 0.1, 0.2$ . Here we suppose  $c_2 = 1$  in (11). From Table 7, we find that the three methods produce the worse estimator as the contamination percentage  $\gamma$  increases. The contaminating data have a more significant impact on the WCLS and MQL methods than the CLS method. The explanation for this phenomenon may be that the WCLS and MQL methods use more wrong information in the contaminating data due to the existence of the weight functions. However, the WCLS and MQL methods are still better than the CLS method especially when we consider the SD. As before, the WCLS and MQL methods give the competitive results and the WCLS method is a little worse than the MQL method when the contaminating data exists.

In the two simulation studies, we find that the WCLS and MQL methods can produce more satisfactory results than the CLS method. While the WCLS method is reliable in each situation, it may cause inconvenience because choosing a suitable weight is a problem which can not be avoided. Based on the above discussions, we conclude that the inverse of the conditional variance is a more satisfactory weight function and we recommend the use of the MQL method to estimate the model parameters.

## 6 Real data analysis

In this section, we conduct two applications to illustrate the usefulness of the GSCINAR(1) process in explaining underdispersed and overdispersed phenomena. The two data sets, exhibiting underdispersion and overdispersion, are used. We compare our process with some INAR(1) models based on the binomial and negative binomial thinning operators:

- POINAR(1) model (Alzaid and Al-Osh 1987);
- NGINAR(1) model (Ristić et al. 2009);
- ZMGINAR(1) model (Barreto-Souza 2015).

The MQL method is used to estimate the unknown parameters of the fitted models. We assume that  $\{\epsilon_t\}$  in the GSCINAR(1) model is a sequence of iid  $GP(\lambda_\epsilon, \kappa_\epsilon)$  random variables. The moment estimators of  $\lambda_\epsilon, \kappa_\epsilon$  are given by

$$\widehat{\kappa}_\epsilon = 1 - \frac{\widehat{\mu}_\epsilon}{\widehat{\sigma}_\epsilon^2}, \quad \widehat{\lambda}_\epsilon = \widehat{\mu}_\epsilon(1 - \widehat{\kappa}_\epsilon),$$

where  $\widehat{\mu}_\epsilon$  is the MQL estimator of  $\mu_\epsilon$  and  $\widehat{\sigma}_\epsilon^2$  is the moment estimator (see (10)). Also, the following statistics of the fitted models are computed: mean, variance, index of dispersion  $I_d$  (the variance to mean ratio), first-order autocorrelation coefficient ACF(1), root mean square of differences between observations and predicted values (RMS) and zero probability  $p_0$ . As before, the zero probability is computed from the average percentages of zeros in time series of length 5000 generated from the corresponding model and the average is obtained based on 1000 replications.



**Table 8** Descriptive statistics for the criminal mischief count data

Mean	Variance	$I_d$	ACF(1)	$p_0$
18.9416	43.2760	2.2847	0.2527	0

**Table 9** Estimators of the parameters and statistics for the criminal mischief count data

Model	Estimators	Mean	Variance	$I_d$	ACF(1)	RMS	$p_0$
GSCINAR(1)	$\hat{\alpha} = -1.9375$						
	$\hat{\lambda}_\epsilon = 8.7781$	19.0005	43.2949	2.2786	0.2713	6.3515	0
	$\hat{\kappa}_\epsilon = 0.3660$						
POINAR(1)	$\hat{\alpha} = 0.2735$	19.0403	19.0403	1.0000	0.2735	6.3516	0
	$\hat{\lambda} = 13.8036$						
NGINAR(1)	$\hat{\alpha} = 0.2681$	18.9997	379.989	19.9997	0.2681	6.3515	0.05
	$\hat{\mu} = 18.9997$						

### 6.1 Modelling overdispersion

In this section, one real example is applied to show good performance of the GSCINAR(1) model in fitting overdispersed count data. We applied our model to fit the series of a monthly count of criminal mischief reported in the twentieth police car beat in Pittsburgh. The data consists of 137 observations starting in January 1990 and ending in May 2001.

Table 8 displays some descriptive statistics of the criminal mischief counts. We find that the sample mean is smaller than the sample variance. Thus, the data set seems to be overdispersed. The zero probability equals to zero indicates that the data set is zero-truncated. A time series plot, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are shown in Fig. 4, which indicates that an autoregressive process of order one is suitable to model the series. In Table 9, we provide the estimators, mean, variance,  $I_d$ , ACF(1), RMS and  $p_0$  of the fitted models. From the results presented in Table 9, although the POINAR(1) model can capture the zero-truncated characteristic of the data set, it performs worst when we consider ACF(1) and RMS. Furthermore, it is very clear that the POINAR(1) process is not suitable for modelling this data set since it can not explain overdispersion. While the NGINAR(1) model gives the best fit of ACF(1), it fails to describe the overdispersed phenomenon accurately. The GSCINAR(1) model can capture overdispersion accurately and  $I_d$  of the GSCINAR(1) model is very close to the empirical  $I_d$ . Moreover, the GSCINAR(1) model performs well when we consider ACF(1). During our study, we also applied the ZMGINAR(1) model to fit this data set. However, the constraints on the model parameters lead to the result that the ZMGINAR(1) model is not suitable for the data. Based on these facts, we recommend the use of the GSCINAR(1) model to fit this data set.

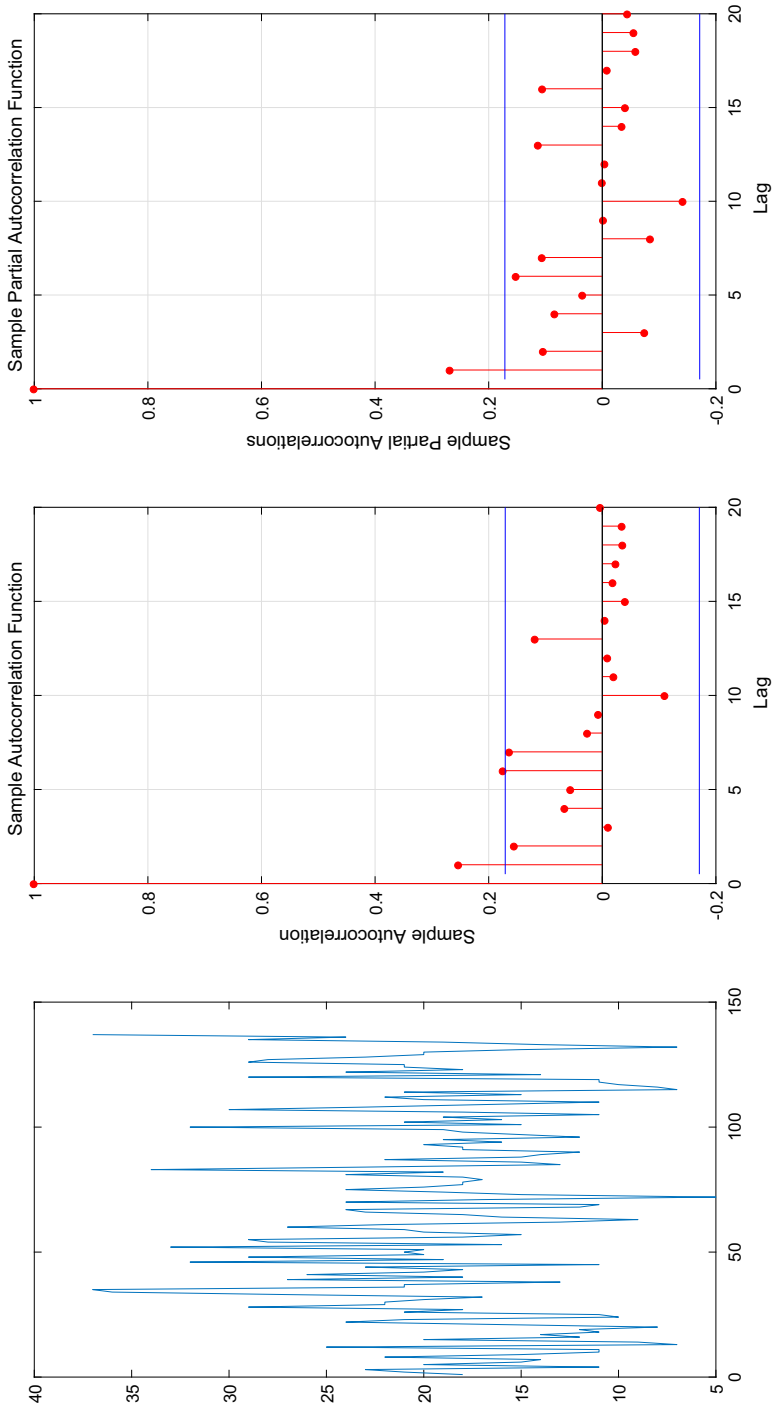


Fig. 4 The sample path, ACF and PACF of the criminal mischief count data

**Table 10** Descriptive statistics for the ip count data

Mean	Variance	$I_d$	ACF(1)	$p_0$
1.2863	1.2052	0.9369	0.2925	0.2739

**Table 11** Estimators of the parameters and statistics for the ip count data

Model	Estimators	Mean	Variance	$I_d$	ACF(1)	RMS	$p_0$
GSCINAR(1)	$\hat{\alpha} = -1.9136$						
	$\hat{\lambda}_\epsilon = 1.0635$	1.2917	1.2070	0.9344	0.2774	1.0464	0.2551
	$\hat{\kappa}_\epsilon = -0.1395$						
POINAR(1)	$\hat{\alpha} = 0.2852$	1.2863	1.2863	1	0.2852	1.0463	0.2748
	$\hat{\lambda} = 0.9232$						
NGINAR(1)	$\hat{\alpha} = 0.2872$	1.2917	2.9601	2.2917	0.2872	1.0463	0.4364
	$\hat{\mu} = 1.2917$						
	$\hat{\pi} = -1.1186$						
ZMGINAR(1)	$\hat{\alpha} = 0.2737$	1.2917	1.1983	0.9277	0.2737	1.0465	0.1976
	$\hat{\mu} = 0.6097$						

## 6.2 Modelling underdispersion

To illustrate the usefulness of the GSCINAR(1) process in modelling underdispersion, we consider an observation of this time series corresponds to the number of different IP addresses ( $\approx$  different users) registered within periods of 2-min length at the server of the Department of Statistics of the University of Würzburg in November and December 2005. In particular, we focus on the time series collected on November 29th, 2005, between 10 o'clock in the morning and 6 o'clock in the evening, a time series of length 241. These data have been investigated by Weiß (2007, 2008a) and Zhu (2012a, b).

Some descriptive statistics of the data are shown in Table 10, which reveals that the data set shows underdispersion since the empirical  $I_d$  is smaller than one. The plots of the data, ACF and PACF are presented in Fig. 5. Within these fitted models, the estimators, mean, variance,  $I_d$ , ACF(1), RMS and  $p_0$  are summarized in Table 11, which shows that the NGINAR(1) model is not a good choice since it gives the wrong information that the data set is overdispersion. The POINAR(1) model has the best results when we consider some statistics. To be specific, it gives the best fit when we consider the mean, ACF(1), RMS and  $p_0$ . However, the GSCINAR(1) model gives the most satisfactory results among the alternative models based on the variance and  $I_d$ . It is well-known that the POINAR(1) model encounters the problem that it can only deal with equidispersion. For this data set, the POINAR(1) model fails to account for the underdispersed phenomenon. The ZMGINAR(1) model is practicable in this case, since  $\hat{\pi} = -1.1186$  and  $\hat{\mu} = 0.6097$  satisfy the constraint  $\pi \in (-1/\mu, 1)$ . The ZMGINAR(1) and GSCINAR(1) models capture the underdispersed feature well and the GSCINAR(1) model is a little better than the ZMGINAR(1) model when we consider  $I_d$ . Based on ACF(1), RMS and  $p_0$ , the GSCINAR(1) model also gives the better fit than the ZMGINAR(1) model. In summary, the GSCINAR(1) process gives

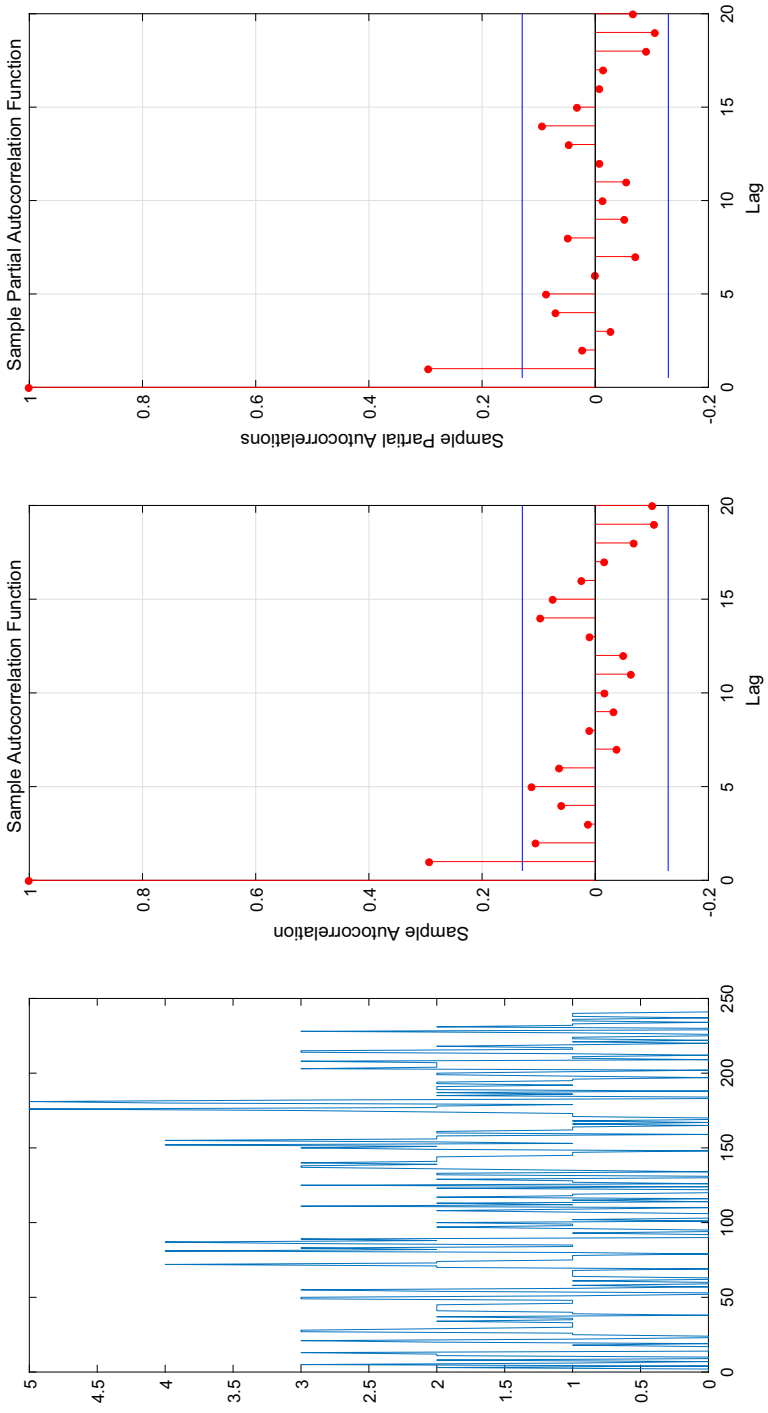


Fig. 5 The sample path, ACF and PACF of the ip count data

the satisfactory fits based on each statistics and the most comprehensive performances among the alternative models. We conclude that the GSCINAR(1) model is the best choice for fitting this data set.

### 7 Discussion

In this paper, we have introduced GSCINAR(1) process. The strict stationarity, ergodicity and some statistical properties of the process are obtained. The CLS, WCLS and MQL methods are used to estimate the model parameters. Two real examples show that our model not only can model the underdispersed data but also has the ability to explain the overdispersed phenomenon.

However, more research is still necessary for some aspects of the GSCINAR(1) process. One of the most important issues may be extending the GSCINAR(1) process to the higher-order autoregressive model.

**Definition 3** An INAR( $p$ ) model based on the GSC thinning operator, denoted by GSCINAR( $p$ ), is defined by the following difference equation:

$$X_t = \alpha_1 \diamond X_{t-1} + \alpha_2 \diamond X_{t-2} + \dots + \alpha_p \diamond X_{t-p} + \epsilon_t, \quad t \geq 1, \tag{12}$$

where  $\alpha_i \diamond X_{t-i} = \sum_{j=1}^{X_{t-i}} W_j, i = 1, \dots, p, \{W_j\}$  is a sequence of iid  $GSC(\alpha_i, \exp\{-|\alpha_i|\})$  random variables with the finite mean  $\phi_i$  and variance  $\beta_i, \alpha_i < 1, \alpha_i \neq 0$ . Here, we write  $\phi_i = \frac{1}{\log(1-\alpha_i)} \sum_{s=1}^{\infty} \log(1 - \alpha_i \exp\{-s|\alpha_i|\})$  and  $\beta_i = \frac{1}{\log(1-\alpha_i)} \sum_{s=1}^{\infty} (2s - 1) \log(1 - \alpha_i \exp\{-s|\alpha_i|\}) - \phi_i^2$ .  $\{\epsilon_t\}$  is an innovation sequence of iid non-negative integer-valued random variables, uncorrelated with the past values of  $\{X_t\}$ . Let  $\mu_\epsilon = E(\epsilon_t), \sigma_\epsilon^2 = \text{Var}(\epsilon_t)$  (we assume that they exist).

We give some statistical properties of the GSCINAR( $p$ ) model in the following proposition. The proof of the proposition is similar to the proof of Proposition 2.1 in Zhang et al. (2010) and we omit it.

**Proposition 2** Suppose  $\{X_t\}$  is a stationary process satisfying (12). Then for  $t \geq 1$ ,

- (i)  $E(X_t|X_{t-i}, i = 1, \dots, p) = \sum_{i=1}^p \phi_i X_{t-i} + \mu_\epsilon$ ,
- (ii)  $E(X_t) = \mu_\epsilon / (1 - \sum_{i=1}^p \phi_i)$ ,
- (iii)  $\text{Var}(X_t|X_{t-i}, i = 1, \dots, p) = \sum_{i=1}^p \phi_i X_{t-i} + \sigma_\epsilon^2$ ,
- (iv)  $\text{Var}(X_t) = [\mu_\epsilon \sum_{i=1}^p \phi_i + \sigma_\epsilon^2 (1 - \sum_{i=1}^p \phi_i)] / [(1 - \sum_{i=1}^p \phi_i)(1 - \sum_{i=1}^p \phi_i^2)]$ ,
- (v)  $\rho_k = \text{Corr}(X_{t+k}, X_t) = \sum_{i=1}^p \phi_i \rho_{k-i}, k = 1, \dots$ ,

where  $\phi_i$  and  $\beta_i$  are given in Definition 3.

The strict stationarity and ergodicity of the GSCINAR( $p$ ) model are given by the following theorem. Again, the proof is omitted because it is similar to the proof of Theorem 2.1 in Zhang et al. (2010).

**Theorem 5** *If all roots of the polynomial  $\lambda^p - \phi_1\lambda^{p-1} - \dots - \phi_{p-1}\lambda - \phi_p = 0$  are inside the unit circle, then there exists a unique strictly stationary integer-valued random series  $\{X_t\}$  satisfying*

$$X_t = \alpha_1 \diamond X_{t-1} + \alpha_2 \diamond X_{t-2} + \dots + \alpha_p \diamond X_{t-p} + \epsilon_t, \quad t \geq 1,$$

$\text{Cov}(X_s, \epsilon_t) = 0$  for  $s < t$ . Furthermore, the process is an ergodic process.

We must point out that the GSCINAR( $p$ ) model deserves a more detailed analysis in a future research. In particular, the topic of parameter estimation should be treated in more detail. For example, it would be interesting in applying the empirical likelihood approach to the GSCINAR( $p$ ) model and investigating the asymptotic behavior of the estimators. Furthermore, the forecasting problem for the GSCINAR( $p$ ) model would be particularly relevant for practice.

**Acknowledgements** We gratefully acknowledge the associate editor and anonymous reviewers for their serious work and thoughtful suggestions that have helped us improve this paper substantially. This work is supported by National Natural Science Foundation of China (Nos. 11731015, 11571051, 11501241, 11871028), Natural Science Foundation of Jilin Province (Nos. 20150520053JH, 20170101057JC, 20180101216JC), Program for Changbaishan Scholars of Jilin Province (2015010), and Science and Technology Program of Jilin Educational Department during the “13th Five-Year” Plan Period (No. 2016316).

### Appendix

As we mentioned in the third paragraph of Sect. 2, the infinite sum in the mean and variance for GSC( $\alpha, \theta$ ) are convergent, where  $\alpha < 1, \alpha \neq 0$  and  $0 < \theta < 1$ . We illustrate it, as follows:

Let  $\alpha < 0$ . Denote  $S_n = \sum_{s=1}^n \log(1 - \alpha\theta^s)$ . Then, we have

$$0 \leq \lim_{n \rightarrow \infty} \sup_{p > 0} |S_{n+p} - S_n| = \lim_{n \rightarrow \infty} \sup_{p > 0} \sum_{s=n+1}^{n+p} \log(1 - \alpha\theta^s) \leq \lim_{n \rightarrow \infty} \sup_{p > 0} [p \log(1 - \alpha\theta^n)] = 0.$$

By the Cauchy criterion of series, the infinite sum  $\sum_{s=1}^{\infty} \log(1 - \alpha\theta^s)$  are convergent.

Denote  $S'_n = \sum_{s=1}^n (2s - 1) \log(1 - \alpha\theta^s)$ . Then, we have

$$\begin{aligned} 0 &\leq \lim_{n \rightarrow \infty} \sup_{p > 0} |S'_{n+p} - S'_n| = \lim_{n \rightarrow \infty} \sup_{p > 0} \sum_{s=n+1}^{n+p} (2s - 1) \log(1 - \alpha\theta^s) \\ &\leq \lim_{n \rightarrow \infty} \sup_{p > 0} p \cdot 2(n + p) \log(1 - \alpha\theta^{n+1}) \leq \lim_{n \rightarrow \infty} \sup_{p > 0} 2p(n + p)(-\alpha)\theta^{n+1} = 0, \end{aligned}$$

using  $x \geq \log(1 + x)$  for  $x \geq 0$ . By the Cauchy criterion of series, the infinite sum  $\sum_{s=1}^{\infty} (2s - 1) \log(1 - \alpha\theta^s)$  are convergent. Following the same way, we can see that the two infinite sum are convergent when  $0 < \alpha < 1$ . □

**Proof of Proposition 1** We have (i) and (iii), i.e.,

$$E(X_t|X_{t-1}) = E(\alpha \diamond X_{t-1} + \epsilon_t|X_{t-1}) = \phi X_{t-1} + \mu_\epsilon$$

and

$$\begin{aligned} \text{Var}(X_t|X_{t-1}) &= \text{Var}(\alpha \diamond X_{t-1} + \epsilon_t|X_{t-1}) \\ &= \text{Var}(\alpha \diamond X_{t-1}|X_{t-1}) + \text{Var}(\epsilon_t|X_{t-1}) \\ &= \beta X_{t-1} + \sigma_\epsilon^2. \end{aligned}$$

Then, we get

$$E(X_t) = E[E(X_t|X_{t-1})] = \phi E(X_{t-1}) + \mu_\epsilon$$

and

$$\begin{aligned} \text{Var}(X_t) &= \text{Var}[E(X_t|X_{t-1})] + E[\text{Var}(X_t|X_{t-1})] \\ &= \text{Var}(\phi X_{t-1} + \mu_\epsilon) + E(\beta X_{t-1} + \sigma_\epsilon^2) \\ &= \phi^2 \text{Var}(X_{t-1}) + \beta E(X_{t-1}) + \sigma_\epsilon^2, \end{aligned}$$

which yield (ii) and (iv), due to the stationarity;  $E(X_t) = E(X_{t-1})$  and  $\text{Var}(X_t) = \text{Var}(X_{t-1})$ . Moreover, we have (v), i.e.,

$$\begin{aligned} \text{Cov}(X_t, X_{t+k}) &= \text{Cov}(X_t, \underbrace{\alpha \diamond \dots \diamond \alpha}_k \diamond X_t) + \text{Cov}(X_t, \sum_{j=0}^{k-1} \underbrace{\alpha \diamond \dots \diamond \alpha}_j \diamond \epsilon_{t+k-j}) \\ &= E\{E[X_t(\underbrace{\alpha \diamond \dots \diamond \alpha}_k \diamond X_t)|X_t]\} - E(\underbrace{\alpha \diamond \dots \diamond \alpha}_k \diamond X_t) \cdot E(X_t) \\ &= \phi^k \{E(X_t^2) - [E(X_t)]^2\} \\ &= \phi^k \text{Var}(X_t). \end{aligned}$$

□

**Proof of Theorem 1** We first introduce a random sequence  $\{X_t^{(n)}\}$ ,

$$X_t^{(n)} = \begin{cases} 0, & n < 0, \\ \epsilon_t, & n = 0, \\ \alpha \diamond X_{t-1}^{(n-1)} + \epsilon_t, & n > 0, \end{cases}$$

where  $\text{Cov}(X_s^{(n)}, \epsilon_t) = 0$  when  $s < t$  for any  $n$ .

As in Li et al. (2015), we can verify: existence of  $\{X_t\}$  satisfying (5), i.e., (A1)  $X_t^{(n)} \in L^2, n > 0$ , (A2)  $X_t^{(n)}$  is a Cauchy sequence, (A3)  $\{X_t\}$  satisfies (5), uniqueness, strict stationarity and ergodicity. The details are omitted here to save space. □

**Proof of Theorem 2** From (6), solving  $\partial Q_1(\eta)/\partial\alpha = 0$  and  $\partial Q_1(\eta)/\partial\mu_\epsilon = 0$  lead to the CLS estimators of  $\alpha$  and  $\mu_\epsilon$ . Now, let  $\mathcal{F}_n = \sigma\{X_0, X_1, \dots, X_n\}$ ,  $M_n^{(1)} = -\frac{1}{2}(\partial Q_1(\eta)/\partial\alpha) = \sum_{t=1}^n \dot{\phi}X_{t-1}(X_t - \phi X_{t-1} - \mu_\epsilon)$ ,  $M_0^{(1)} = 0$ . Also,  $M_n^{(2)} = -\frac{1}{2}(\partial Q_1(\eta)/\partial\mu_\epsilon) = \sum_{t=1}^n (X_t - \phi X_{t-1} - \mu_\epsilon)$ ,  $M_0^{(2)} = 0$ . Then, it is easy to see that  $\{M_n^{(1)}, \mathcal{F}_n\}_{n \geq 0}$  and  $\{M_n^{(2)}, \mathcal{F}_n\}_{n \geq 0}$  are martingales. The martingale central limit theorem and Cramer-Wold’s device imply that

$$n^{-1/2}(M_n^{(1)}, M_n^{(2)})' \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{CLS}).$$

Using Taylor’s expansion, we have

$$\mathbf{0} = -\frac{1}{2\sqrt{n}} \frac{\partial Q_1(\widehat{\eta}_{CLS})}{\partial \eta} = -\frac{1}{2\sqrt{n}} \frac{\partial Q_1(\eta)}{\partial \eta} - \frac{1}{2n} \frac{\partial^2 Q_1(\eta)}{\partial \eta \partial \eta'} \sqrt{n}(\widehat{\eta}_{CLS} - \eta) + o_p(n^{-1/2}).$$

Since we have proved that  $-\frac{1}{2\sqrt{n}} \frac{\partial Q_1(\eta)}{\partial \eta} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{CLS})$ , after some algebra, we have

$$\sqrt{n}(\widehat{\eta}_{CLS} - \eta) \xrightarrow{d} N(\mathbf{0}, \mathbf{H}_{CLS}^{-1} \mathbf{V}_{CLS} \mathbf{H}_{CLS}^{-1}).$$

This completes the proof. □

**Proof of Theorem 4** Following Zheng et al. (2007), we firstly suppose  $\tau$  is known. Let

$$L_n^{(1)}(\tau, \eta) = \sum_{t=1}^n V_\tau^{-1}(X_t|X_{t-1}) \dot{\phi}X_{t-1}(X_t - \phi X_{t-1} - \mu_\epsilon), \quad L_0^{(1)}(\tau, \eta) = 0,$$

$$L_n^{(2)}(\tau, \eta) = \sum_{t=1}^n V_\tau^{-1}(X_t|X_{t-1})(X_t - \phi X_{t-1} - \mu_\epsilon), \quad L_0^{(2)}(\tau, \eta) = 0.$$

Similar to Theorem 2, we have

$$n^{-1/2}(L_n^{(1)}(\tau, \eta), L_n^{(2)}(\tau, \eta))' \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{MQL}).$$

Now, we replace  $V_\tau^{-2}(X_t|X_{t-1})$  by  $V_{\widehat{\tau}}^{-2}(X_t|X_{t-1})$ , where  $\widehat{\tau}$  is a consistent estimator of  $\tau$ . Then we want

$$n^{-1/2}(L_n^{(1)}(\widehat{\tau}, \eta), L_n^{(2)}(\widehat{\tau}, \eta))' \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{MQL}).$$

For this we need to prove that  $\frac{1}{\sqrt{n}}L_n^{(i)}(\widehat{\tau}, \eta) - \frac{1}{\sqrt{n}}L_n^{(i)}(\tau, \eta) \xrightarrow{P} 0$ ,  $i = 1, 2$  [its proof is omitted here, since the argument is the same as in Zheng et al. (2007)]. Following the proof of Theorem 2, by Taylor’s expansion and some algebra, we have

$$\sqrt{n}(\widehat{\eta}_{MQL} - \eta) \xrightarrow{d} N(\mathbf{0}, \mathbf{H}_{MQL}^{-1} \mathbf{V}_{MQL} \mathbf{H}_{MQL}^{-1}).$$



This completes the proof. □

## References

- Alzaid, A. A., & Al-Osh, M. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8, 261–275.
- Barreto-Souza, W. (2015). Zero-modified geometric INAR(1) process for modelling count time series with deflation or inflation of zeros. *Journal of Time Series Analysis*, 36, 839–852.
- Barreto-Souza, W. (2017). Mixed Poisson INAR(1) processes. *Statistical Papers*, <https://doi.org/10.1007/s00362-017-0912-x>. (in press).
- Borges, P., Molinara, F. F., & Bourguignon, M. (2016). A geometric time series model with inflated-parameter Bernoulli counting series. *Statistics and Probability Letters*, 119, 264–272.
- Bourguignon, M., Rodrigues, J., & Santosneto, K. (2019). Extended Poisson INAR(1) processes with equidispersion, underdispersion and overdispersion. *Journal of Applied Statistics*, 46, 101–118.
- Bourguignon, M., & Vasconcellos, K. L. (2015). First order non-negative integer valued autoregressive processes with power series innovations. *Brazilian Journal of Probability and Statistics*, 29, 71–93.
- Bourguignon, M., & Weiß, C. H. (2017). An INAR(1) process for modeling count time series with equidispersion, underdispersion and overdispersion. *Test*, 26, 847–868.
- Cossette, H., Marceau, É., & Tourelle, F. (2011). Risk models based on time series for count random variables. *Insurance: Mathematics and Economics*, 48, 19–28.
- Freeland, R. K., & McCabe, B. P. M. (2004). Analysis of low count time series data by Poisson autoregression. *Journal of Time Series Analysis*, 25, 701–722.
- Gómez-Déniz, E., Sarabia, J. M., & Calderín-Ojeda, E. (2011). A new discrete distribution with actuarial applications. *Insurance: Mathematics and Economics*, 48, 406–412.
- Grunwald, G., Hyndman, R. J., Tedesco, L., & Tweedie, R. L. (2000). Non-Gaussian conditional linear AR(1) models. *Australian and New Zealand Journal of Statistics*, 42, 479–495.
- Jazi, M. A., Jones, G., & Lai, C. D. (2012). First-order integer valued AR processes with zero inflated poisson innovations. *Journal of Time Series Analysis*, 33, 954–963.
- Karlsen, H., & Tjøstheim, D. (1988). Consistent estimates for the Near(2) and Nlar(2) time series models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 50, 313–320.
- Kim, H. Y., & Lee, S. (2017). On first-order integer-valued autoregressive process with Katz family innovations. *Journal of Statistical Computation and Simulation*, 87, 546–562.
- Li, C., Wang, D., & Zhang, H. (2015). First-order mixed integer-valued autoregressive processes with zero-inflated generalized power series innovations. *Journal of the Korean Statistical Society*, 44, 232–246.
- Ristić, M. M., Bakouch, H. S., & Nastić, A. S. (2009). A new geometric first-order integer-valued autoregressive (NGINAR(1)) process. *Journal of Statistical Planning and Inference*, 139, 2218–2226.
- Schweer, S., & Weiß, C. H. (2014). Compound Poisson INAR (1) processes: Stochastic properties and testing for overdispersion. *Computational Statistics and Data Analysis*, 77, 267–284.
- Scotto, M. G., Weiß, C. H., & Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: A review. *Statistical Modelling*, 15, 590–618.
- Steutel, F. W., & Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7, 893–899.
- Weiß, C. H. (2007). Controlling correlated processes of Poisson counts. *Quality and Reliability Engineering International*, 23, 741–754.
- Weiß, C. H. (2008a). Serial dependence and regression of Poisson INARMA models. *Journal of Statistical Planning and Inference*, 138, 2975–2990.
- Weiß, C. H. (2008b). Thinning operations for modeling time series of counts—a survey. *ASTA Advances in Statistical Analysis*, 92, 319–341.
- Weiß, C. H., Homburg, A., & Puig, P. (2019). Testing for zero inflation and overdispersion in INAR(1) models. *Statistical Papers*, 60, 473–498.
- Zhang, H., Wang, D., & Zhu, F. (2010). Inference for INAR( $p$ ) processes with signed generalized power series thinning operator. *Journal of Statistical Planning and Inference*, 140, 667–683.
- Zheng, H., Basawa, I. V., & Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference*, 137, 212–229.

- Zhu, F. (2012a). Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *Journal of Mathematical Analysis and Applications*, 389, 58–71.
- Zhu, F. (2012b). Modeling time series of counts with COM-Poisson INGARCH models. *Mathematical and Computer Modelling*, 56, 191–203.

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