Journal of the Korean Statistical Society 44 (2015) 232–246

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

First-order mixed integer-valued autoregressive processes with zero-inflated generalized power series innovations

Cong Li, Dehui Wang *, Haixiang Zhang

School of Mathematics, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 11 January 2014
Accepted 30 August 2014
Available online 29 September 2014

AMS 2000 subject classifications:
primary 62M10
secondary 62P99

Keywords:
Binomial thinning
Generalized power series distribution
Negative binomial thinning
Zero inflation
ZIMINAR(1) process

ABSTRACT

To model zero-inflated time series of counts, we propose a first-order mixed integer-valued autoregressive process with zero-inflated generalized power series innovations. These innovations contain the commonly used zero-inflated Poisson and geometric distributions. Strict stationarity, ergodicity of the process, and some important probabilistic properties such as the transition probabilities, the k-step ahead conditional mean and variance are obtained. The conditional maximum likelihood estimators for the parameters in this process are derived and the performances of the estimators are studied via simulation. As illustration, an application to an offence data set is given to show the effectiveness of the proposed model.

1. Introduction

Count data with extra zeros, which is also known as zero inflation, is commonly seen in many fields, such as econometrics, manufacturing defects, medical consultations, and sexual behaviour (Ridout, Demétrio, & Hinde, 1998). There are many cases that will lead to zero inflation. For example, when counting the abundance of rare species, the extra zeros can be considered coming from two possible sources: 'stochastic' zeros from a distribution of counts where the conditions are suitable for animal presence at a site, and structural zeros where animal presence is not tenable (Welsh, Cunningham, Donnelly, & Lindenmayer, 1996). Extra zeros also lead to overdispersion, which violates the variance–mean relationship of the usually assumed Poisson structure. So count data with many zeros should be assessed in both zero inflation and overdispersion. However, many researchers may not account for zero inflation, instead, they simply use a negative binomial model to deal with these data. Via simulation, Perumean-Chaney, Morgan, McDowall, and Aban (2013) pointed out that when zero inflation in the data was ignored, the estimation results were poor and some potentially significant statistic findings were missed, the misspecifications caused by the ignorance of zero inflation may even lead to erroneous conclusions about the data and bring uncertainty to the research and application.

Zero-inflated Poisson and negative binomial models, hurdle models, semi-parametric hurdle models, birth process models and threshold models are the usual regression models for zero-inflated count data and have been reviewed and compared on application potentials in agricultural and horticultural research by Ridout et al. (1998). In general, the zero-inflated model can be viewed as a mixture of a degenerate distribution with mass at zero and a non degenerate distribution such as the Poisson or negative binomial distribution. Until now, there are many authors concentrating on the construct and selection

* Corresponding author.
E-mail address: wangdh@jlu.edu.cn (D. Wang).

http://dx.doi.org/10.1016/j.jkss.2014.08.004
1226-3192/

of regression models, as well as the estimation of the regression parameters of zero-inflated count data (Fox, 2013; Ospina & Ferrari, 2012; Staub & Winkelmann, 2013). However, count data with so many zeros may exhibit strong autocorrelation. Take the offence data reported in the 47st police car beat in Pittsburgh as an example, the data consists of 144 observations, starting in January 1990 and ending in December 2001. The histogram given in Fig. 5 shows a substantial number of zeros (approximately 70%), the autocorrelation function (ACF) and partial autocorrelation function (PACF) obtained in Figs. 6 and 7 reveal that the data may fit a first-order autoregressive process. This type of count data time series has received sustained attention, see Al-Osh and Alzaid (1987), Alzaid and Al-Osh (1988, 1993), Du and Li (1991), Latour (1998), and Zhu and Joe (2006). Many new results on integer-valued time series have been obtained further in recent years. Ristić, Bakouch, and Nastić (2009) constructed a new stationary first-order integer-valued autoregressive (INAR) process with geometric marginal distributions based on negative binomial thinning. Zhang, Wang, and Zhu (2010) proposed a pth-order integer-valued autoregressive processes with signed generalized power series thinning operator. Nastić and Ristić (2012) considered some mixed integer-valued autoregressive models of orders 1 and 2 with geometric marginal distributions. Ristić and Nastić (2012) introduced a mixed INAR(p) model. Jazi, Jones, and Lai (2012) derived a new stationary first-order integer valued autoregressive process with zero-inflated Poisson innovations. Zhu (2012) gave zero-inflated Poisson and negative binomial integer-valued GARCH models. Wang and Zhang (2011) and Zhang, Wang, and Zhu (2012) proposed a series of generalized random coefficient integer-valued autoregressive processes with signed thinning operator. Some semiparametric INAR models where there are essentially no restrictions on the innovation distribution are also considered. Drost, Akker, and Werker (2009) provided an efficient estimator of a semiparametric INAR(p) model. McCabe, Martin, and Harris (2011) derived probabilistic forecasts of a semiparametric INAR(p) model. Meanwhile, Fokianos (2011) reviewed some regression models for the analysis of count time series. Tjøstheim (2012) gave an overview on recent theoretical developments in autoregressive count time series.

The mixed integer-valued autoregressive time series is an important research field, which is a probabilistic mixture of binomial and negative binomial thinning operators. The binomial thinning operator based on a counting series of Bernoulli distributed random variables is appropriate for modelling the number of random events, which may only survive or vanish after a period of observation. As to these realistic phenomena in which counted events may also become potentially more active, generating new random events, the negative binomial thinning operator based on a counting series of geometric distributed random variables becomes very appropriate instead. So, if we want to deal with the situation when the "activity" rate of the observed elements is changing from one time interval to another in the process, we should combine these two thinning operators. In this paper, to model the above described situation and analyse zero-inflated autocorrelated integer-valued time series, we propose a mixed first-order integer-valued autoregressive process with zero-inflated generalized power series innovations.

The paper is organized as follows. In Section 2, we introduce a first-order mixed integer-valued autoregressive processes with zero-inflated generalized power series innovations, denoted by ZIMINAR(1). The strict stationarity and ergodicity of the proposed model are established. In Section 3, several important probabilistic properties such as the transition probabilities, the k-step ahead conditional mean and variance are derived. In Section 4, estimators of the model parameters are obtained by using the conditional maximum likelihood method. Some simulation results for the estimators are presented. In Section 5, we present an application of the proposed model to a real offence data. Finally, we conclude in Section 6 and outline issues for future research.

2. Construction of the ZIMINAR(1) process

In this section, we introduce a new process called ZIMINAR(1) to handle the non-negative integer-valued time series with zero inflation. Let X be a discrete random variable with probability distribution function

$$P(X = k) = \frac{a(k)(g(\theta))^k}{f(\theta)}, \quad k = 0, 1, 2, \ldots, \tag{2.1}$$

where $a(k) > 0$, $g(\theta)$ and $f(\theta)$ are positive, finite, differentiable functions, then we call this a generalized power series distribution. Now, let us recall the definition of zero-inflated generalized power series distribution (ZIGPSD), which will be used to construct the new ZIMINAR(1) process.

Definition 1. A discrete random variable Y is said to have a zero-inflated generalized power series distribution if Y is a mixture of the Dirac measure at 0 and generalized power series distribution X, i.e., with probability distribution function

$$P(Y = k) = \phi I_{\{0\}}(k) + (1 - \phi)P(X = k), \quad k = 0, 1, 2, \ldots, \tag{2.2}$$

where $0 \leq \phi \leq 1$, $I_{\{0\}}(k) = 1$ for $k = 0$ and 0 else; the distribution of X is defined in (2.1).

Remark 1. (i) The ZIGPSD includes among others the zero-inflated generalized Poisson and zero-inflated generalized negative binomial distributions and hence the zero-inflated Poisson, zero-inflated binomial and zero-inflated negative binomial distributions.

(ii) Let $\mu_1 = (1/f(\theta)) \sum_x xa(x)(g(\theta))^x$, then from Gupta, Gupta, and Tripathi (1995), we have

$$E(X) = (1 - \phi)\frac{g(\theta)}{g'(\theta)}\frac{d}{d\theta}\ln f(\theta), \qquad \mathrm{Var}(X) = (1 - \phi)\frac{g(\theta)}{g'(\theta)}\frac{d\mu_1}{d\theta} + \phi(1 - \phi)\left(\frac{g(\theta)}{g'(\theta)}\frac{d}{d\theta}\ln f(\theta)\right)^2.$$