Journal of Statistical Planning and Inference 142 (2012) 826–839

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Zero-inflated Poisson and negative binomial integer-valued GARCH models

Fukang Zhu

School of Mathematics, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 18 January 2011
Received in revised form
7 July 2011
Accepted 10 October 2011
Available online 15 October 2011

Keywords:
EM algorithm
Integer-valued GARCH model
Negative binomial
Poisson
Time series of counts
Zero inflation

ABSTRACT

Zero inflation means that the proportion of 0's of a model is greater than the proportion of 0's of the corresponding Poisson model, which is a common phenomenon in count data. To model the zero-inflated characteristic of time series of counts, we propose zero-inflated Poisson and negative binomial INGARCH models, which are useful and flexible generalizations of the Poisson and negative binomial INGARCH models, respectively. The stationarity conditions and the autocorrelation function are given. Based on the EM algorithm, the estimating procedure is simple and easy to be implemented. A simulation study shows that the estimation method is accurate and reliable as long as the sample size is reasonably large. A real data example leads to superior performance of the proposed models compared with other competitive models in the literature.

1. Introduction

In the probability model the Poisson distribution is usually assumed for count data; however, in many real applications it is likely to observe that the number of zeroes is greater than what would be expected for the Poisson model, which is called zero inflation. The zero inflation is of interest because zero counts frequently have special status, e.g., in counting disease lesions on plants, a plant may have no lesions either because it is resistant to the disease, or simply because no disease spores have landed on it. This is the distinction between *structural zeros*, which are inevitable, and *sampling zeros*, which occur by chance (Ridout et al., 1998). Ignoring zero inflation can have at least two consequences; first, the estimated parameters and standard errors may be biased, and second, the excessive number of zeros can cause overdispersion (Zuur et al., 2009, p. 269).

In recent years there has been considerable and growing interest in modeling zero-inflated count data, and many models have been proposed, e.g., the hurdle model (Mullahy, 1986), the zero-inflated Poisson (ZIP) model (Lambert, 1992), and the two-part model (Heilbron, 1994, also known as the zero-altered model). Ridout et al. (1998) reviewed this literature and cited examples from econometrics, manufacturing defects, patent applications, road safety, species abundance, medical consultations, use of recreational facilities, and sexual behavior. For the ZIP model, Böhning (1998) also reviewed the related literature and provided a variety of examples from different disciplines. As a generalization of the ZIP model, the zero-inflated negative binomial (ZINB) model has been discussed by many authors, such as Ridout et al. (2001) considered the score test for testing the ZIP model against the ZINB model. Zeileis et al. (2008) gave a nice overview and comparison of Poisson, negative binomial, and zero-inflated models in the software R. For a recent review and applications to ecology, see Zuur et al. (2009, Chapter 11).

E-mail address: zfk8010@163.com

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.10.002

In general, zero-inflated model can be viewed as a mixture of a degenerate distribution with mass at zero and a nondegenerate distribution such as the Poisson or negative binomial distribution. Now many researchers are still studying that how to extend and test these models (see, e.g., Xie et al., 2009; Yang et al., 2009; Min and Czado, 2010; Hall and Shen, 2010; Garay et al., 2011). To our best knowledge, all the zero-inflated models are considered in regression context, not yet in time series context, except that Bakouch and Ristić (2010) considered a zero-truncated Poisson INAR(1) model. But zero inflation is also common in time series analysis, see the example given in Section 6.

In addition to the zero-inflated characteristic, many time series count datasets also display overdispersion, which means that the variance is greater than the mean. Overdispersion has been well modeled in the literature, such as the integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) model proposed by Ferland et al. (2006) and its various generalizations. The INGARCH model is defined as follows:

$$\begin{cases} X_t|\mathcal{F}_{t-1} : \mathcal{P}(\lambda_t), \quad \forall t \in \mathbb{Z}, \\ \lambda_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}, \end{cases} \tag{1.1}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$, and $\mathcal{F}_{t-1}$ is the σ-field generated by $\{X_{t-1}, X_{t-2} \ldots\}$. This model has been studied by many authors. Zhu et al. (2008), Zhu and Li (2009) and Zhu and Wang (2010, 2011) considered various estimation and testing methods. Specially, Zhu and Wang (2011) gave a necessary and sufficient condition for the existence of higher-order moments. Fokianos et al. (2009) considered geometric ergodicity and likelihood-based inference, and Fokianos and Fried (2010) transferred the concept of intervention effects to model (1.1). Weiß (2009) derived a set of equations from which the variance and the autocorrelation function of the general case can be obtained. Weiß (2010a) derived the unconditional distributions via the Poisson–Charlier expansion, while Weiß (2010b) considered higher-order moments and jumps. Zhu et al. (2010) extended model (1.1) to the mixture model context, while Zhu (2011) extended the Poisson deviate to the negative binomial one, which are useful generalizations. For more generalizations, see Fokianos and Tjøstheim (2011) and Matteson et al. (2011). Fokianos (2011) reviewed some recent progress in INGARCH models.

To model overdispersion and zero inflation in the same framework, we will generalize the Poisson model (1.1) and the negative binomial model proposed in Zhu (2011) and show the usefulness of these generalizations. The paper is organized as follows. In Sections 2 and 3 we describe the zero-inflated Poisson INGARCH (ZIP-INGARCH) model and the zero-inflated negative binomial INGARCH (ZINB-INGARCH) model, respectively. The stationarity conditions and the autocorrelation functions are given. We discuss the estimation procedure in Section 4 via the EM algorithm. Section 5 presents a simulation study. In Section 6 we apply the proposed models to a real data example. Section 7 gives some discussions.

2. The zero-inflated Poisson INGARCH model

First, recall the definition of ZIP distribution (see Johnson et al., 2005, Section 4.10.3). A distribution is said to be ZIP (λ, ω) if its probability mass function (pmf) can be written in the form

$$P(X = k) = \omega \delta_{k,0} + (1-\omega)\frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2 \ldots,$$

where $0 < \omega < 1$, $\delta_{k,0}$ is the Kronecker delta, i.e., $\delta_{k,0}$ is 1 when $k=0$ and is zero when $k \neq 0$. The probability generating function (pgf) is $G(z) = \omega + (1-\omega)e^{\lambda(z-1)}$, then from Lemma 1 in Ferland et al. (2006) we know that the uncentered moments of X satisfy

$$E(X^m) = (1-\omega) \sum_{j=0}^{m} \mathfrak{S}_m^{(j)} \lambda^j, \tag{2.1}$$

where $\mathfrak{S}_m^{(j)}$ is the Stirling number of the second kind (for details, see Gradshteyn and Ryzhik, 2007, p. 1046). Specially, we have

$$E(X) = (1-\omega)\lambda, \quad \mathrm{Var}(X) = (1-\omega)\lambda(1+\omega\lambda) > E(X).$$

Let $\{X_t\}$ be a time series of counts. We assume that, conditional on $\mathcal{F}_{t-1}$, the random variables $X_1, \ldots, X_n$ are independent, and the conditional distribution of X_t is specified by a ZIP distribution. To be specific, we consider the following model:

$$X_t|\mathcal{F}_{t-1} : \mathcal{ZIP}(\lambda_t, \omega), \quad \lambda_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}, \tag{2.2}$$

where $0 < \omega < 1$, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$, $\mathcal{F}_{t-1}$ is the σ-field generated by $\{X_{t-1}, X_{t-2} \ldots\}$. The above model is denoted by ZIP-INGARCH(p, q). The conditional mean and conditional variance of X_t are given by

$$E(X_t|\mathcal{F}_{t-1}) = (1-\omega)\lambda_t, \quad \mathrm{Var}(X_t|\mathcal{F}_{t-1}) = (1-\omega)\lambda_t(1+\omega\lambda_t), \tag{2.3}$$