Long-memory models for count time series
Modelli a memoria lunga per serie di dati di conteggio

Luisa Bisaglia, Massimiliano Caporin and Matteo Grigoletto

Abstract In this paper we analyze two different approaches for modeling dependent count data with long-memory. The first model we consider explicitly takes into account the integer nature of data and the long-range correlation, while the second model is a count-data long-memory model where the distribution of the current observation is specified conditionally upon past observations. We compare these two different models by looking at their estimation and forecasting performances.
Abstract *In questo lavoro analizziamo due diversi approcci per modellare la dipendenza di lungo periodo in serie storiche di dati di conteggio. Il primo modello considera esplicitamente la natura dei dati e la correlazione di lungo periodo, mentre il secondo è un modello a memoria lunga per dati di conteggio in cui la distribuzione dell'osservazione attuale viene specificata condizionatamente alle osservazioni passate. Il confronto fra questi due approcci è fatto tramite uno stidio Monte Carlo che confronta le performance di stima e previsione.*

Key words: count time series, long-memory, GLM, estimation, forecasting

1 Introduction

Recently, there has been a growing interest in studying nonnegative integer-valued time series and, in particular, time series of counts. In some cases, the discrete values of the time series are large numbers and may be analyzed using continuous-valued models such as ARMA with Gaussian errors. However, when the values are small,

L. Bisaglia
Dept. Statistical Sciences, University of Padova, e-mail: luisa.bisaglia@unipd.it

M. Caporin
Dept. Statistical Sciences, University of Padova, e-mail: massimiliano.caporin@unipd.it

M. Grigoletto
Dept. Statistical Sciences, University of Padova e-mail: matteo.grigoletto@unipd.it

as in the case of counting processes, the usual linear ARMA processes become inappropriate for modeling and forecasting purposes since they would invariably produce non-integer forecast values. One of the most common approaches to build an integer-valued autoregressive (INAR) process is based on a probabilistic operator called binomial thinning, as reported in Al-Osh and Alzaid (1987) and McKenzie (1985) who first introduced INAR processes. A different approach is based on the generalized linear models (GLM) advanced by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). This framework generalizes the traditional ARMA methodology allowing for more flexible dynamics also coherent with count data time series (for details and references, see, for example, Fokianos, 2016).

Long-memory (LM) processes have proved to be useful tools in the analysis of many empirical time series. One of the most popular processes that takes into account this particular behavior of the autocorrelation function is the AutoRegressive Fractionally Integrated Moving Average process (ARFIMA(p,d,q)), independently introduced by Granger and Joyeux (1980) and Hosking (1981). This process generalizes the ARIMA(p,d,q) process by relaxing the assumption that d is an integer. In particular, when $d \in (0,0.5)$ the autocorrelation function of the process decays to zero hyperbolically at a rate $O(k^{2d-1})$, where k denotes the lag. If $p = q = 0$, the process $\{X_t,\, t = 0, \pm 1, \ldots\}$ is called Fractionally Integrated Noise, FI(d). In the following we will concentrate on FI(d) processes with $d \in (0,0.5)$.

Persistent count time series occur for example in finance when modeling stock market daily trading volumes (e.g. Palma and Zevallos, 2011).
In this work, we analyze two different approaches for modeling dependent count data with long-memory. The first model we consider takes explicitly into account the integer nature of data and the long-range correlation, mixing the INteger AutoRegressive (INAR) model proposed by Al-Osh and Alzaid (1987) and McKenzie (1985) with the Fractionally Integrated (FI) model introduced by Granger and Joyeux (1980) and Hosking (1981). The second model, introduced by Palma and Zevallos (2011), builds on a conditional distribution for count data where the parameters' dynamic is characterized by long-memory. We compare estimation and forecasting performances of the two models by Monte Carlo simulations.

2 LM models for count time series data

2.1 GLM approach

Palma and Zevallos (2011) introduce a model for count time series characterized by long-range dependence. They propose a new class of conditional long-memory models (CLMs), where the conditional distribution of the data, given a data-driven parameter, is explicitly specified. The conditional long-memory process, X_t, can be defined as follows:

$$X_t \mid \mathscr{F}_{t-1} \sim G(\lambda_t, g(\lambda_t)), \quad \text{with} \quad \lambda_t = \mu \sum_{j=0}^{\infty} \pi_j - \sum_{j=1}^{\infty} \pi_j X_{t-j} \tag{1}$$

where $\mathscr{F}_t$ is the $\sigma-$field generated by the information up to instant t, $\{X_t, X_{t-1}, \ldots\}$, $G(\alpha, \beta)$ is a distribution corresponding to a continuous or discrete nonnegative random variable with mean α and variance β, both finite, $g(\cdot)$ is a positive function, μ is a constant and $\{\pi_j\}_{j>0}$ is an absolutely summable sequence of real numbers such that $\pi_0 = 1$ and $\pi_j \approx C j^{-d-1}$ for large positive j and some $d < 0.5$. Therefore, given the information $\mathscr{F}_{t-1}$, X_t has distribution G with conditional mean λ_t and conditional variance $g(\lambda_t)$. Obviously, if G is a distribution corresponding to a discrete nonnegative random variable, we obtain a LM model for count time series data. Even if, from a theoretical point of view, this setup is general enough to allow for the use of different integer distributions, in practice only a Poisson has been used by Palma and Zevallos (2011).

It can be shown (Palma and Zevallos, 2011) that model (1) has a non-Gaussian ARFIMA(p,d,q) representation. In particular, for the ARFIMA$(0,d,0)$ case we have $\pi_0 = 1$ and $\pi_j = \Gamma(j-d)/[\Gamma(j+1)\Gamma(-d)]$, for $j \geq 1$.

For further details about CLM-ARFIMA processes see Palma and Zevallos (2011). In particular, it is shown that CLM-ARFIMA and standard ARFIMA processes share the same correlation structure.

The CLM approach allows using all the tools available for GLM models (see, for instance, Liboschik et al., 2017).

2.2 Models based on the thinning operator

Integer-valued autoregressive (INAR) processes, initially proposed by Al-Osh and Alzaid (1987) and McKenzie (1985), in their most basic form follow the recursion:

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t$$

where '$\circ$'is the thinning operator, defined by $\alpha \circ X = \sum_{i=1}^{X} Y_i$ with $X \in \mathbb{N}$, $\alpha \in [0,1]$ and Y_i is a sequence of i.i.d. count random variables, typically Ber(α), independent of X with common mean α. Hence, α plays the role of thinning probability. Moreover, ε_t is a sequence of i.i.d. discrete random variables with mean μ_ε and variance σ_ε^2. While the INAR(1) and INMA(1) models are defined univocally, for the INAR(p) and INMA(q) models there are additional complexities and different types of INAR(p) and INMA(q) processes might be considered (see Weiss, 2018 for a recent review on this topic). Recently, Weiss (2019) developed the INARMA(p,q) process:

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + \beta_1 \circ \varepsilon_{t-1} + \ldots + \beta_q \circ \varepsilon_{t-q} + \varepsilon_t$$

where, to obtain feasible stochastic properties, he assumed independence among all thinnings, independence from the innovations, and independence from $(X_s)_{s<t}$

for the thinning at time t. Combining the ideas of the INARMA model with the fractional integration of Granger and Joyeux (1980) and Hosking (1981), Quoreshi (2014) introduce the INARFIMA model based on the following INMA(∞) representation:

$$X_t = \sum_{i=0}^{\infty} \psi_i \circ \varepsilon_{t-i} \tag{2}$$

where $\psi_0 = 1$ and $\psi_i = \Gamma(i+d)/[\Gamma(i+1)\Gamma(d)]$, for $i \geq 1$, and d is the long memory coefficient. Since the ψ_i in (2) are considered thinning probabilities, then $d \in [0,1]$. Quoreshi (2014) proposes different estimation methods, based on conditional least squares, feasible generalized least squares and the generalized method of moments. In his paper Quoreshi (2014) does not consider the problem of forecasting with the INARFIMA model.

In the present paper, differently from Quoreshi (2014), who adopts the MA(∞) representation, to take into account the long memory and integer nature of data, we propose to consider the INAR(∞) recursion:

$$X_t = \sum_{i=1}^{\infty} \pi_i \circ X_{t-i} + \varepsilon_t \tag{3}$$

that is:

$$\sum_{i=0}^{\infty} \pi_i \circ X_{t-i} = (1 - B^{\circ})^d X_t = \varepsilon_t \tag{4}$$

where $\pi_0 = 1$ and $\pi_j = \Gamma(j-d)/[\Gamma(j+1)\Gamma(-d)]$ for $j \geq 1$, with $d \in (0, 0.5)$. As in Du and Li (1991) the ε_t constitutes a sequence of i.i.d. discrete random variables independent of all counting series, and all thinning operations are mutually independent. The conditional mean of process (3) is given by:

$$E[X_t \mid X_{t-1}, \ldots] = \mu_\varepsilon + \sum_{i=1}^{\infty} \pi_i X_{t-i}$$

and thus the autocorrelation function of X_t is the same of an I(d) process. Moreover, the conditional variance is:

$$V[X_t \mid X_{t-1}, \ldots] = \sigma_\varepsilon^2 + \sum_{i=1}^{\infty} \pi_i (1 - \pi_i) X_{t-i} \ .$$

In practice, only $X_1, \ldots, X_n$, are available, but X_t depends on the infinite past of the process. Therefore, we must approximate X_t with an AR(p), taking p large enough so that $\sum_{i=p+1}^{\infty} \pi_i \circ X_{t-i}$ in (3) is negligible. In our applications, all the available past observations are used.

2.3 Forecasting LM models for count time series

For CLM-ARFIMA models the natural one-step predictor of λ_{n+1} conditional on the past information, $\mathscr{F}_n$, is based on (1) and can be written as: $\hat{\lambda}_{n+1} = \hat{\mu}\sum_{j=0}^{n}\hat{\pi}_j - \sum_{j=1}^{n}\hat{\pi}_j X_{n+1-j}$ where each $\hat{\pi}_j$ depends on the long-memory parameter estimate, $\hat{d}$ (and other parameters, if present). Hence, the predicted conditional distribution is

$$\hat{X}_{n+1} \mid \mathscr{F}_n \sim G(\hat{\lambda}_{n+1}, g(\hat{\lambda}_{n+1}))$$

and the construction of conditional prediction intervals for one-step forecasts is a simple task. For k-step forecasts, with $k > 1$, we have to recursively use previous forecast values, for example: $\hat{\lambda}_{n+2} = \hat{\mu}\sum_{j=0}^{n+1}\hat{\pi}_j - \hat{\pi}_{n+1}\hat{X}_{n+1} - \sum_{j=2}^{n+1}\hat{\pi}_j X_{n+2-j}$ and the predicted conditional distribution is

$$\hat{X}_{n+2} \mid \mathscr{F}_n \sim G(\hat{\lambda}_{n+2}, g(\hat{\lambda}_{n+2}))\ .$$

In this case, the construction of conditional prediction intervals is not immediate and we obtain prediction intervals via computational methods. Forecasting with INARFIMA models is very simple too. Using (3), we have:

$$\hat{X}_{n+k} = \sum_{i=1}^{\infty}\hat{\pi}_i \circ \hat{X}_{n+k-i}\ ,$$

with $\hat{X}_j = X_j$ for $j \leq n$. Also in this case, prediction intervals cannot be directly recovered and we must resort to computational methods.

3 Some simulation results

In this section we provide the results of some Monte Carlo experiments we carried out to asses the estimation performance of different long-memory parameter estimators. Count time series of lengths $n = 500$ and $n = 1000$ were generated from models (1) and (3). The Poisson and Negative Binomial distributions were used for the conditional distribution G in (1) (models CP and CNB) and for ε_t in the INAR model (3) (models PI and NBI). The distribution parameters were chosen so that the amount of over-dispersion in the different cases is comparable. The functions we use are written in the `R` language and are available upon request from the authors. Table 1 shows the average estimates (and their standard deviations, in parentheses) over 1000 Monte Carlo replications. The estimation methods considered are maximum likelihood (ML), Geweke and Porter-Hudak (GPH) and Whittle (WH). Simulation results show how the long memory parameter d is, for all models, correctly estimated on average, with the ML and WH methods yielding comparable standard deviations, while the GPH method performs considerably worse.

		NBI				CNB			
d	n	OD	ML	GPH	WH	OD	ML	GPH	WH
0.15	500	1.5	0.14 (0.035)	0.16 (0.169)	0.14 (0.036)	1.1	0.14 (0.036)	0.16 (0.165)	0.14 (0.037)
	1000		0.15 (0.026)	0.15 (0.136)	0.15 (0.026)		0.14 (0.026)	0.15 (0.132)	0.15 (0.026)
0.35	500	1.4	0.34 (0.037)	0.36 (0.168)	0.34 (0.039)	1.5	0.34 (0.037)	0.36 (0.168)	0.34 (0.039)
	1000		0.35 (0.025)	0.39 (0.141)	0.35 (0.026)		0.34 (0.026)	0.35 (0.141)	0.35 (0.027)
0.45	500	2.9	0.43 (0.030)	0.47 (0.181)	0.44 (0.034)	2.2	0.43 (0.033)	0.46 (0.169)	0.44 (0.036)
	1000		0.44 (0.024)	0.48 (0.141)	0.45 (0.027)		0.44 (0.024)	0.46 (0.144)	0.45 (0.027)
		PI				CP			
d	n	OD	ML	GPH	WH	OD	ML	GPH	WH
0.15	500	5.5	0.14 (0.038)	0.15 (0.165)	0.14 (0.038)	1.1	0.14 (0.037)	0.15 (0.169)	0.14 (0.038)
	1000		0.15 (0.026)	0.16 (0.137)	0.15 (0.026)		0.15 (0.025)	0.15 (0.135)	0.15 (0.025)
0.35	500	2.5	0.34 (0.035)	0.37 (0.166)	0.35 (0.036)	1.4	0.34 (0.036)	0.35 (0.182)	0.34 (0.038)
	1000		0.35 (0.026)	0.38 (0.146)	0.35 (0.027)		0.34 (0.025)	0.35 (0.138)	0.35 (0.026)
0.45	500	3.8	0.43 (0.030)	0.48 (0.166)	0.45 (0.034)	2.7	0.43 (0.031)	0.47 (0.175)	0.44 (0.035)
	1000		0.44 (0.022)	0.48 (0.145)	0.45 (0.025)		0.44 (0.023)	0.47 (0.131)	0.45 (0.025)

Table 1 Estimation of the long memory parameter d for series generated from different models, having comparable over-dispersions (OD). The considered models are the Poisson INAR (PI), Conditional Poisson (CP), Negative Binomial INAR (NBI) and Conditional Negative Binomial (CNB). The estimation methods are maximum likelihood (ML), the Geweke and Porter-Hudak estimator (GPH) and the Whittle estimator (WH). Results show the average estimates and their standard deviations (in parentheses) over 1000 Monte Carlo replications.

References

1. Al-Osh, M. A. and A. A. Alzaid: First order integer-valued autoregressive INAR(1) process. Journal of Time Series Analysis **8**, 261–275 (1987)
2. Du, J.G. and Y. Li: The integer-valued autoregressive INAR(p) model. Journal of Time Series Analysis **12**, 129–142 (1991)
3. Fokianos, K.: Statistical analysis of count time series models: A GLM perspective. In Davis R., Holan S., Lund R., and Ravishanker R. (eds.), Handbook of Discrete-Valued Time Series, pp. 3-27. Chapman & Hall/CRC (2016)
4. Granger, C. and R. Joyeux: An introduction to long-range time series models and fractional differencing. Journal of Time Series Analysis **1**, 15–30 (1980)
5. Hosking, J.: Fractional differencing. Biometrika **68**, 165–176 (1981)
6. Liboschik, T., K. Fokianos, and F. R.: tscount: An R package for analysis of count time series following generalized linear models. Journal of Statistical Software **82**, 1–51 (2017)
7. McCullagh, P. and J. Nelder: Generalized Linear Models. London, U.K. Chapman & Hall (1989)
8. McKenzie, E.: Some simple models for discrete variate time series. Water Resources Bulletin **21**, 645–650 (1985)
9. Nelder, J. and R. Wedderburn: Generalized Linear Models. Journal of the Royal Statistical Society, Series A, **135**, 370–384 (1972)
10. Palma, W. and M. Zevallos: Fitting non-Gaussian persistent data. Applied Stochastic Models in Business and Industry **27**, 23–36 (2011)
11. Quoreshi, A: A long-memory integer-valued time series model, INARFIMA, for financial application. Quantitative Finance **14**, 2225–2235 (2014)