Journal of Multivariate Analysis 102 (2011) 563–578

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Log-linear Poisson autoregression

Konstantinos Fokianos [a,*], Dag Tjøstheim [b]

[a] Department of Mathematics & Statistics, University of Cyprus, Cyprus

[b] Department of Mathematics, University of Bergen, Norway

ARTICLE INFO

Article history:
Received 13 April 2010
Available online 25 November 2010

AMS subject classifications:
primary 62M10
secondary 62F12

Keywords:
Autocorrelation
Covariates
Ergodicity
Generalized linear models
Perturbation
Prediction
Stationarity
Volatility

ABSTRACT

We consider a log-linear model for time series of counts. This type of model provides a framework where both negative and positive association can be taken into account. In addition time dependent covariates are accommodated in a straightforward way. We study its probabilistic properties and maximum likelihood estimation. It is shown that a perturbed version of the process is geometrically ergodic, and, under some conditions, it approaches the non-perturbed version. In addition, it is proved that the maximum likelihood estimator of the vector of unknown parameters is asymptotically normal with a covariance matrix that can be consistently estimated. The results are based on minimal assumptions and can be extended to the case of log-linear regression with continuous exogenous variables. The theory is applied to aggregated financial transaction time series. In particular, we discover positive association between the number of transactions and the volatility process of a certain stock.

1. Introduction

Let $\{Y_t\}$ be a time series of counts and $\mathcal{F}_t^{Y,\lambda}$ the σ-field generated by $\{Y_0, \ldots, Y_t, \lambda_0\}$, that is $\mathcal{F}_t^{Y,\lambda} = \sigma(Y_s, s \leq t, \lambda_0)$, where $\{\lambda_t\}$ is a Poisson intensity process. In a recent contribution, Fokianos et al. [10,11] studied in detail the following autoregressive model for $\{Y_t\}$:

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + a\lambda_{t-1} + bY_{t-1},\ t \geq 1, \tag{1}$$

where the parameters d, a, b are assumed to be positive and satisfy $0 < a + b < 1$. Related work concerning the properties of (1) has been carried out by Rydberg and Shephard [23], Streett [24] and Ferland et al. [8]. As it turns out, when $0 < a + b < 1$, then there exist initial distributions for Y_0 and λ_0 such that the process $\{Y_t\}$ is stationary with mean $\mathrm{E}[Y_t] = \mathrm{E}[\lambda_t] \equiv \mu = d/(1 - a - b)$ and autocovariance function

$$\mathrm{Cov}[Y_t, Y_{t+h}] = \begin{cases} \dfrac{(1-(a+b)^2+b^2)\mu}{1-(a+b)^2}, & h = 0, \\[2ex] \dfrac{b(1-a(a+b))(a+b)^{h-1}\mu}{1-(a+b)^2}, & h \geq 1. \end{cases}$$

* Corresponding author.
E-mail addresses: fokianos@ucy.ac.cy (K. Fokianos), Dag.Tjostheim@math.uib.no (D. Tjøstheim).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2010.11.002

Although model (1) appears to provide an adequate framework for modeling count dependent data, there are at least two drawbacks related to its application. We first note that $\mathrm{Cov}[Y_t, Y_{t+h}] > 0$ because $0 < a + b < 1$. Therefore model (1) cannot be employed for modeling negative correlation. However alternative count time series models have been suggested for resolving this particular issue; see [16,6], for instance. An additional shortcoming of (1) is that it can include exclusively covariates which result in a positive regression term, since otherwise the mean of the Poisson process becomes negative.

The purpose of this paper is to enrich the class of autoregressive models for count time series by providing alternative structures which accommodate both positive and negative association and, even more importantly, can include covariates in a straightforward manner. The point of view that we take towards this goal is based on the generalized linear model (GLM) theory; see [19] for modeling independent data and [15] for time series data. More precisely, if $\{Y_t\}$ is a count time series then we will be working with the so-called canonical link process $\{\nu_t\}$, which is defined by $\nu_t \equiv \log \lambda_t$. Suppose that $\mathcal{F}_t^{Y,\nu}$ denotes the σ-field generated by $\{Y_0, \ldots, Y_t, \nu_0\}$. We study the following family of log-linear autoregressive models:

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\nu} \sim \text{Poisson}(\lambda_t), \qquad \nu_t = d + a\nu_{t-1} + b\log(Y_{t-1} + 1), \quad t \geq 1. \tag{2}$$

The parameters d, a, b belong to $\mathbb{R}$ but restrictions on the parameter space will be imposed so that a central limit theory for $\{(Y_t, \nu_t)\}$ can be developed. Both ν_0 and Y_0 are assumed to be fixed. Note that the log-intensity process of (2) is expressed as

$$\nu_t = d\frac{1 - a^t}{1 - a} + a^t \nu_0 + b\sum_{i=0}^{t-1} a^i \log(1 + Y_{t-i-1}), \tag{3}$$

after repeated substitution. Therefore, the hidden process ν_t is determined by past functions of lagged responses. Hence, (2) belongs to the class of observation driven models in the sense of Cox [5].

Arguments in favor of this model will be given throughout the article but the essence of (2) is that it takes into account both negative and positive correlation and it can include real-valued covariates. In addition, model (2) is expected to be more parsimonious than a model which includes higher lags of $\log(Y_t + 1)$ but without the feedback mechanism introduced by ν_t. This is analogous to GARCH models (see [1]) being more parsimonious than the ARCH models.

We choose to work with a log-linear model where lagged observations of the response Y_t are fed into the autoregressive equation for ν_t via the term $\log(Y_{t-1} + 1)$. This is a one-to-one transformation of Y_{t-1} which is quite standard in coping with zero data values and it maps zeros of Y_{t-1} into zeros of $\log(Y_{t-1} + 1)$. Moreover, both λ_t and Y_t are transformed onto the same scale. Covariates can be accommodated by including them in the second equation of (2). An alternative modeling approach is based upon employing the transformation $\log(\max(Y_{t-1}, c))$ (cf. [28]) for $c \in (0, 1]$, instead of $\log(Y_{t-1} + 1)$ in (2). Certainly other specifications are possible—for a short discussion see Section 4 which specifies a model for the log-mean process by introducing $\log(Y_{t-1} + \upsilon)$, where υ is a constant varying from 1 to 10 with a step equal to 0.5. The results of the data analysis do not indicate any gross deviations in terms of the mean square error of residuals from a model that includes $\log(Y_{t-1} + 1)$.

To motivate further the choice of the $\log(\cdot)$ function for the lagged values of the response, consider a model like (2) but with Y_{t-1} included instead of $\log(Y_{t-1} + 1)$. In other words, set

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\nu} \sim \text{Poisson}(\lambda_t), \qquad \nu_t = d + a\nu_{t-1} + bY_{t-1}.$$

But then

$$\lambda_t = \exp(d)\lambda_{t-1}^a \exp(bY_{t-1}),$$

and therefore stability of the above system is guaranteed only when $b < 0$. Otherwise, the process λ_t increases exponentially fast; see [27] and [15, Ch.4] for more. Hence, only negative correlation can be introduced by such a model. However (2) allows for positive (respectively, negative) correlation by allowing the parameter b to take positive (respectively, negative) values.

Log-linear models for time series of counts have been considered by several authors, including Zeger and Qaqish [28], Li [17], MacDonald and Zucchini [18], Brumback et al. [3], Kedem and Fokianos [15], Davis et al. [7], Fokianos and Kedem [9] and Jung et al. [14]. Model (2) is related to the work of Zeger and Qaqish [28] and Li [17] but these authors did not address either the problem of ergodicity or the problem of asymptotic inference. We point out that Zeger and Qaqish [28] heuristically argued that the restriction $b < 1$ is a sufficient condition for stability of (2), provided that $a = 0$. In fact, this condition comes out as a special case of Proposition 2.1. The work of Davis et al. [7] addresses theoretical properties for the following specification of the log-mean process:

$$\nu_t = \beta_0 + \beta_1\frac{Y_{t-1} - \lambda_{t-1}}{\lambda_{t-1}^{\alpha}} = \beta_0 + \beta_1\frac{(Y_{t-1} - \exp(\nu_{t-1}))}{\exp(\alpha\nu_{t-1})},$$

where β_0 and β_1 ($\beta_1 \neq 0$) are regression parameters and $\alpha \in (0, 1]$. It was shown by Davis et al. [7] that if $1/2 \leq \alpha \leq 1$ then the chain $\{\nu_t\}$ has a stationary distribution. In particular, when $\alpha = 1$, then $\{\nu_t\}$ is uniformly ergodic and has a unique stationary distribution.

This paper is a follow-up to Fokianos et al. [10,11]. Some of the results of the present paper can be obtained using techniques similar to those developed in these papers. For such results, we just briefly indicate the derivations, referring the reader to Fokianos et al. [10,11] for details. This is the case for the likelihood theory of Section 3. On the other hand, the stability conditions for the present model are far more complex and not straightforward to derive. The results are found in Section 2 with proofs given in the Appendix. Also, considerable emphasis is put on the data example and inclusion of covariates in Section 4, since these aspects illustrate much wider potential applications of model (2).

Table 1
Autocorrelation function at lag 1 derived from model (2) for selected values of the parameters a and b when $d = 0.5$. Results are based on 10 000 data points.

a	−0.800	−0.500	−0.400	0.100	0.250	0.250
b	−0.430	−1.000	−0.350	0.200	0.550	0.730
$\rho(1)$	−0.979	−0.500	−0.202	0.150	0.637	0.980

2. Structure

2.1. Preliminaries

To investigate empirically the behavior of processes such as (2), we resort initially to simulation. Fig. 1 shows time series plots of 200 observations from (2) and their associated sample autocorrelation function, for different parameter configurations. For all the plots, $\nu_0 = 1$ is chosen, but different values of the starting value do not affect the results. In fact these observations are taken after generating 500 data points and discarding the first 300 observations in order to make sure that we are in the stationary region; see Proposition 2.1 and Lemma 2.1.

It is a rather challenging problem to get an explicit expression for the autocorrelation function of model (2), similar to the one obtained for model (1). The computational difficulty stems from the form of the log-linear representation, since unlike the case for (1), we cannot obtain a recursive relationship for the covariance. However, by simulating a very long path of the series, we get a clue to the range of possible values of the correlation obtained by (2). Table 1 illustrates the autocorrelation function $\rho(\cdot)$ of model (2) at lag 1. Note that the log-linear model can produce both large negative and large positive correlations. The parameter values have been selected according to Proposition 2.1. Hence, our initial claim that (2) produces both positive and negative correlations is verified, at least empirically. Fig. 1(a) indicates that when b is positive but $|a + b| < 1$, then we obtain a time series with a positive lag 1 autocorrelation function but the dependence decays for larger lags. However, when both a and b are positive but their sum is closer to 1, then there is stronger dependence—see Fig. 1(b) which demonstrates that the values of the process $\{Y_t\}$ become large. When both a and b are negative then the lag 1 autocorrelation function of the process is negative.

2.2. Ergodicity

In what follows we obtain sufficient conditions for proving ergodicity of a perturbed version of (2). Towards this goal it is more convenient to express the sequence of independent Poisson drawings in the first part of (2), more explicitly in terms of random variables. Introduce for each time point t a Poisson process $N_t(\cdot)$ of unit intensity. Then, the first equation of (2) can be restated in terms of these Poisson processes by assuming that Y_t, given λ_t, is equal to the number of events $N_t(\lambda_t)$ (equivalently $N_t(\exp(\nu_t))$) of $N_t(\cdot)$ in the time interval $[0, \lambda_t]$ (equivalently $[1, \exp(\nu_t)]$). Let therefore $\{N_t(\cdot), t = 1, 2, \ldots\}$ be a sequence of independent Poisson processes of unit intensity and rephrase (2) as

$$Y_t = N_t(\lambda_t), \qquad \nu_t = d + a\nu_{t-1} + b\log(Y_{t-1} + 1), \tag{4}$$

for $t \geq 1$ and with Y_0, λ_0 fixed. This notation will be used throughout the paper, and it is emphasized that models (2) and (4) are equivalent.

To study the ergodic properties of model (2), we operate along the lines of Fokianos et al. [10,11] who introduce the perturbed chain $\{(Y_t^m, \nu_t^m)\}$ defined by

$$Y_t^m = N_t(\lambda_t^m) = N_t(\exp(\nu_t^m)), \quad \nu_t^m = d + a\nu_{t-1}^m + b\log(Y_{t-1}^m + 1) + \varepsilon_{t,m}, \tag{5}$$

with ν_0^m, Y_0^m fixed, where $\{N_t(\cdot)\}$ is identical to the sequence $\{N_t(\cdot)\}$ of (4), and

$$\varepsilon_{t,m} = c_m 1\left(Y_{t-1}^m = 1\right) U_t, \quad c_m > 0,\ c_m \to 0,\ \text{as } m \to \infty,$$

where $1(\cdot)$ is the indicator function, and where $\{U_t\}$ is a sequence of i.i.d. uniform random variables on $(0, 1)$ such that U_t is independent of $N_t(\cdot)$. It was shown in [10,11] that the inclusion of such a perturbation bypasses several problems related to establishing geometric ergodicity of model (1). Therefore, the method is expected to be useful when applied to models like (2). An interesting alternative approach could possibly be via coupling theory. Actually, Franke [12] has recently used coupling for studying properties of model (1). See also [21] for related work on model (1).

The first step in proving geometric ergodicity is based on ϕ-irreducibility. The problem in attempting to prove ϕ-irreducibility for model (2) is that as Y_t varies over the integers, then $\log(Y_{t-1} + 1)$ also varies over a fixed countable set of numbers, whereas our candidate for a ϕ-irreducible measure is the Lebesgue measure over a certain set. It is this fact that has prompted us to introduce the perturbed system (5). Note that for both (2) and (5), the skeleton of the system is $\nu_t = d + a\nu_{t-1}$. Then $\nu^* = d/(1 - a)$ is a solution of $\nu = d + a\nu$, i.e., a fixed point of the mapping $f(\nu) = d + a\nu$.

The perturbation can be chosen in many other ways; for example $\{U_t\}$ can be an i.i.d. sequence of positive random variables with bounded support and finite moments. It turns out that the likelihood functions for $\{Y_t\}$ and $\{Y_t^m\}$ as far as dependence on $\{\nu_t\}$ is concerned are identical for models (2) and (5). In addition, we will be showing that the geometric

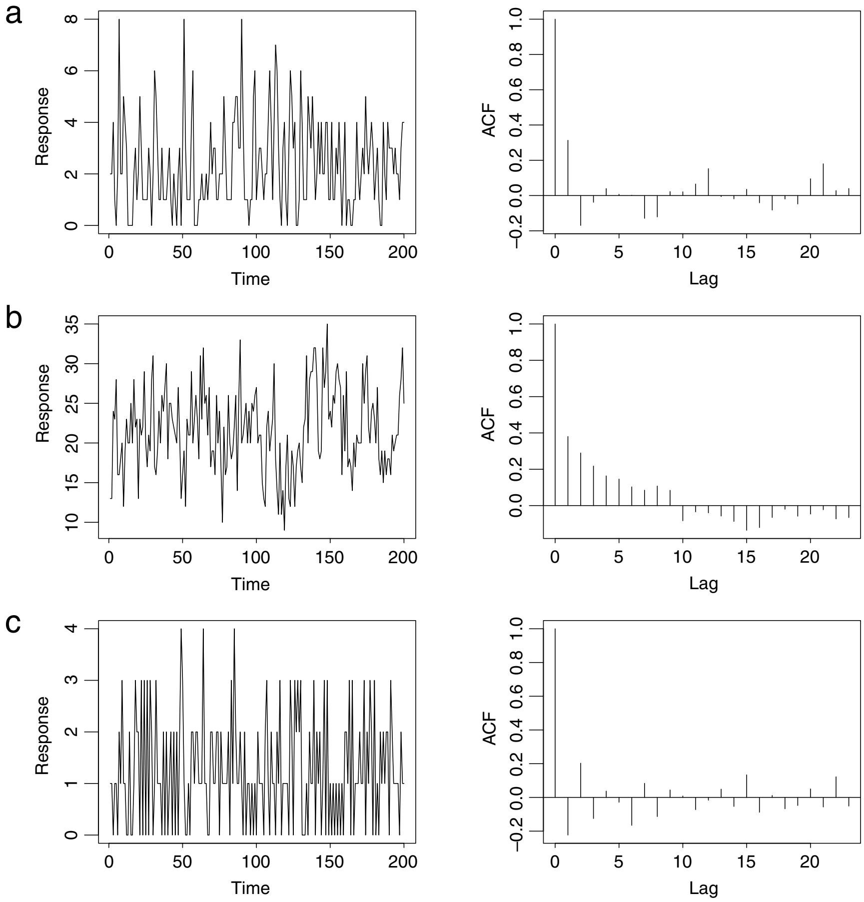

Fig. 1. Two hundred observations and their sample autocorrelation function from model (2) for different parameter values. (a) $d = 0.5$, $a = -0.5$ and $b = 2/3$. (b) $d = 0.5$, $a = 0.5$ and $b = 1/3$. (c) $d = 0.5$, $a = -0.5$ and $b = -1/4$.

ergodicity of the perturbed process yields asymptotic normality for the maximum likelihood estimators. Then, by allowing $c_m \to 0$, we will obtain asymptotic normality of the likelihood estimates of (2). The following results show the more complex stability structure of model (2) as compared to model (1). They also require new proofs that are given in the Appendix. The result implies that the trivariate chain $\{(Y_t^m, U_t, \nu_t^m), t \geq 0\}$ converges to a stationary limit for arbitrary initial conditions and that there exist initial distributions for Y_0^m, U_0 and ν_0^m such that when started with these distributions the trivariate chain would be stationary. It should be noted that the marginal distributions of $\{Y_t^m\}$ are not Poisson distributed.

Proposition 2.1. *Assume model (5) and suppose that $|a| < 1$. In addition, assume that when $b > 0$ then $|a + b| < 1$, and when $b < 0$ then $|a| |a + b| < 1$. Then, the following conclusions hold:*

1. *The process $\{\nu_t^m, t \geq 0\}$ is a geometrically ergodic Markov chain with finite moments of order k, for an arbitrary k.*
2. *The process $\{(Y_t^m, U_t, \nu_t^m), t \geq 0\}$ is a $V_{(Y,U,\nu)}$-geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y, U, \nu) = 1 + \log^{2k}(1 + Y) + \nu^{2k} + U^{2k}$, k being a positive integer.*

The above proposition establishes geometric ergodicity of the joint process (Y_t^m, U_t, ν_t^m); see [20, p. 355]. In addition we obtain results for the moments of functions of (Y_t^m, U_t, ν_t^m) which are useful in deriving bounds for proving asymptotic normality of the maximum likelihood estimators. Note that when $a = 0$, then $b < 1$ is a sufficient condition for ergodicity of the time series. In this case, the chain moves over a fixed set of numbers and the Lebesgue measure is replaced by the counting measure, as a ϕ-measure, in the proof of ergodicity. This particular condition was also argued heuristically in [28] by relating (2) to the theory of branching process.

The following lemma shows that the difference between (2) and (5) can be made negligible as $m \to \infty$ such that $c_m \to 0$. It is proved under the conditions that $|a + b| < 1$ if a and b have the same sign, and $a^2 + b^2 < 1$ if they have different signs. These conditions are quite restrictive when compared to the conditions for geometric ergodicity. It is likely that they can be weakened to at least $|a + b| < 1$ for all possible cases of signs and possibly to the generality of the ergodicity conditions. In many applications it seems that $a > 0$ and $b > 0$ in which case, of course, the above condition is the same as the ergodic one, that is $|a + b| < 1$.

Lemma 2.1. *Suppose that* (Y_t, ν_t) *and* $\left(Y_t^m, \nu_t^m\right)$ *are defined by* (2) *and* (5) *respectively. Assume that* $|a+b| < 1$ *if a and b have the same sign, and* $a^2 + b^2 < 1$ *if a and b have different signs. Then the following statements are true:*

(1) $\mathrm{E}|\nu_t^m - \nu_t| \to 0$ *and* $|\nu_t^m - \nu_t| < \delta_{1,m}$ *almost surely for m large,*
(2) $\mathrm{E}(\nu_t^m - \nu_t)^2 \le \delta_{2,m}$,
(3) $\mathrm{E}|\lambda_t^m - \lambda_t| \le \delta_{3,m}$,
(4) $\mathrm{E}|Y_t^m - Y_t| \le \delta_{4,m}$,
(5) $\mathrm{E}(\lambda_t^m - \lambda_t)^2 \le \delta_{5,m}$,
(6) $\mathrm{E}(Y_t^m - Y_t)^2 \le \delta_{6,m}$,

where $\delta_{i,m} \to 0$ *as* $m \to \infty$ *for* $i = 1, \ldots, 6$. *Furthermore, almost surely, with m large enough,*

$$\left|\lambda_t^m - \lambda_t\right| \le \delta \quad \text{and} \quad \left|Y_t^m - Y_t\right| \le \delta, \quad \text{for any } \delta > 0.$$

3. Likelihood inference

To study likelihood inference for (2) suppose that $\boldsymbol{\theta}$ denotes the three-dimensional vector of unknown parameters, that is $\boldsymbol{\theta} = (d, a, b)'$. The true value of the parameter is denoted by $\boldsymbol{\theta}_0 = (d_0, a_0, b_0)'$. The conditional likelihood function for $\boldsymbol{\theta}$ given the starting value $\lambda_0 = \exp(\nu_0)$ in terms of the observations $Y_1, \ldots, Y_n$ is given by

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t(\boldsymbol{\theta}))\lambda_t^{Y_t}(\boldsymbol{\theta})}{Y_t!}.$$

Hence, the log-likelihood function is given, up to a constant, by

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} \left(Y_t \nu_t(\boldsymbol{\theta}) - \exp(\nu_t(\boldsymbol{\theta}))\right), \tag{6}$$

where $\nu_t(\boldsymbol{\theta}) = d + a\nu_{t-1}(\boldsymbol{\theta}) + b\log(1 + Y_{t-1})$. The score function is defined by

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} (Y_t - \exp(\nu_t(\boldsymbol{\theta}))) \frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{7}$$

where $\partial \nu_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a three-dimensional vector with components given by

$$\frac{\partial \nu_t}{\partial d} = 1 + a\frac{\partial \nu_{t-1}}{\partial d}, \qquad \frac{\partial \nu_t}{\partial a} = \nu_{t-1} + a\frac{\partial \nu_{t-1}}{\partial a}, \qquad \frac{\partial \nu_t}{\partial b} = \log(1 + Y_{t-1}) + a\frac{\partial \nu_{t-1}}{\partial b}. \tag{8}$$

The solution of the equation $\boldsymbol{S}_n(\boldsymbol{\theta}) = 0$, if it exists, yields the conditional maximum likelihood estimator of $\boldsymbol{\theta}$ which is denoted by $\hat{\boldsymbol{\theta}}$. The Hessian matrix for model (2) is obtained from

$$\begin{aligned} \boldsymbol{H}_n(\boldsymbol{\theta}) &= -\sum_{t=1}^{n} \frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \\ &= \sum_{t=1}^{n} \exp(\nu_t((\boldsymbol{\theta}))) \left(\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)' - \sum_{t=1}^{n} (Y_t - \exp(\nu_t(\boldsymbol{\theta}))) \frac{\partial^2 \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}. \end{aligned} \tag{9}$$

We know that Proposition 2.1 guarantees geometric ergodicity of the perturbed model (Y_t^m, ν_t^m). But Lemma 2.1 demonstrates that the difference between ν_t^m and ν_t becomes small, for large m. Therefore we use these results to study the asymptotic properties of the maximum likelihood estimators for process (2). Following Fokianos et al. [10,11] we define the counterparts of expressions (6)–(9) for model (5). The likelihood function, say L^m, including the pseudo-observations $U_1, U_2, \ldots, U_n$, is given by

$$L^m(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t^m(\boldsymbol{\theta}))(\lambda_t^m(\boldsymbol{\theta}))^{Y_t^m}}{Y_t^m!} \prod_{t=1}^{n} f_u(U_t),$$

by the Poisson assumption and the asserted independence of U_t from $(Y_{t-1}^m, \lambda_{t-1}^m)$. Here, $f_u(\cdot)$ denotes the uniform density. Note that $L(\boldsymbol{\theta})$ and $L^m(\boldsymbol{\theta})$ have identical forms with the only exception that (Y_t, ν_t) is replaced by (Y_t^m, ν_t^m). Hence, $\boldsymbol{S}_n^m(\boldsymbol{\theta})$ and $\boldsymbol{H}_n^m(\boldsymbol{\theta})$ have the same form as (7) and (9) – with recursions defined by (8) – but with (Y_t, ν_t) replaced by (Y_t^m, ν_t^m). The solution of the equation $\boldsymbol{S}_n^m(\boldsymbol{\theta}) = 0$ is denoted by $\hat{\boldsymbol{\theta}}^m$.

To study the asymptotic properties of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, for the log-linear model (2), we derive and apply the asymptotic properties of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}^m$ for the perturbed log-linear model (5); see [10,11].

The main tool in linking $\hat{\boldsymbol{\theta}}$ to $\hat{\boldsymbol{\theta}}^m$ is Brockwell and Davis [2, Prop. 6.3.9]. We first show that $\hat{\boldsymbol{\theta}}^m$ is asymptotically normal where for the proof of consistency and asymptotic normality we use the fact that the log-likelihood function is differentiable three times, applying Jensen and Rahbek [13, Lemma 1]. Then we show that the score function, the information matrix and the third derivatives of the perturbed likelihood function tend to the corresponding quantities of the unperturbed likelihood function (6). All these results are stated in the Appendix, for completeness of the presentation.

Introduce lower and upper values of each component of $\boldsymbol{\theta}$, $\delta_L < d_0 < \delta_U$, $-1 < \alpha_L < a_0 < \alpha_U < 1$ and $\beta_L < b_0 < \beta_U$, and in terms of these define

$$O(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} | \delta_L \le d \le \delta_U, -1 < \alpha_L \le a \le \alpha_U < 1 \text{ and } \beta_L \le b \le \beta_U\}. \tag{10}$$

Then the following theorem for $\hat{\boldsymbol{\theta}}$ is true.

Theorem 3.1. *Consider model (2) and suppose that at the true value $\boldsymbol{\theta}_0$, $|a_0 + b_0| < 1$ if a_0 and b_0 have the same sign, and $a_0^2 + b_0^2 < 1$ if a_0 and b_0 have different sign. Then, there exists a fixed open neighborhood $O = O(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ – see (10) – such that with probability tending to 1, as $n \to \infty$, the log-likelihood function (6) has a unique maximum point $\hat{\boldsymbol{\theta}}$. Furthermore, $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal,*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{G}^{-1}),$$

where the matrix $\boldsymbol{G}$ is defined in Lemma A.3. A consistent estimator of $\boldsymbol{G}$ is given by $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})$, where

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \operatorname{Var}\left[\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid \mathcal{F}_{t-1}\right] = \sum_{t=1}^{n} \exp(\nu_t(\boldsymbol{\theta})) \left(\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \nu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'.$$

4. Applications

We illustrate the theory by presenting a simulation study and some real data examples. Maximum likelihood estimators are calculated by direct optimization of the log-likelihood function (6)—details are available from the authors. The starting values of the algorithm are obtained by a routine GLM fit of the model $\nu_t = d + b\log(Y_{t-1} + 1)$. Such models are fitted by iteratively reweighted least squares, as outlined by Kedem and Fokianos [15, pp. 15–16]. Given the estimates of (d, b), say $(\tilde{d}, \tilde{b})$, we initiate the optimization procedure by taking $(\tilde{d}, 0, \tilde{b})$ as the initial value for the unknown parameter vector.

4.1. Simulations

We report a limited simulation study for assessing the performance of the maximum likelihood estimator. Table 2 shows the results of a simulation study for two parameter values: $\boldsymbol{\theta} = (0.50, -0.50, 0.65)'$ and $\boldsymbol{\theta} = (0.50, -0.50, -0.35)'$. The first case corresponds to the condition $a^2 + b^2 < 1$ which yields count time series with positive lag 1 correlation. The other choice of parameter values yields negative lag 1 correlation. In the latter case, the true parameters satisfy the condition $|a+b| < 1$. The third column of the table lists the estimates of the parameters obtained by averaging out the results from all runs. Note that the standard errors correspond to the sampling standard error of the estimates obtained by the simulation. The last three columns of Table 2 show some summary statistics of the sampling distribution of the standardized MLE. When n is large, the asserted asymptotic normality is supported clearly for the first choice of $\boldsymbol{\theta}$. For the second choice of the parameter vector $\boldsymbol{\theta}$, we note that $\hat{a}$ approaches the asserted normality for sample sizes greater than 500. Fig. 2 shows histograms and qq-plots for the sampling distribution of the standardized maximum likelihood estimators for the choice of $\boldsymbol{\theta} = (0.50, -0.50, 0.65)'$. All plots point to the adequacy of the normal approximation.

4.2. Data examples

We illustrate the theory by analyzing data reported for the total number of transactions per minute for the stock Ericsson B for the time period between July 2nd and July 22nd, 2002. We consider the first ten days of reported data – that is 4600 observations – and we aggregate further the series of transactions by considering the total number of trades within two-minute and five-minute intervals. That is, if Y_t denotes the observed number of transactions per minute, then the response time series is defined by the variable $Y_{kt} = \sum_{i=1}^{k} Y_{k(t-1)+i}$, for $k = 2$ (two-minute interval) and $k = 5$ (five-minute interval). Note that Fokianos et al. [10,11] have analyzed the first day of transactions per minute for this particular stock. This analysis employs aggregated data of ten working days, over two-minute and five-minute intervals, respectively.

The data and their sample autocorrelation function are shown in Fig. 3(a)–(b). We only plot the derived series over five-minute intervals but similar results were obtained for the two-minute aggregation scheme. Regardless of the series, the sample mean is always less than their sample variance. In other words, all the series under consideration exhibit overdispersion—a well known feature of count time series.

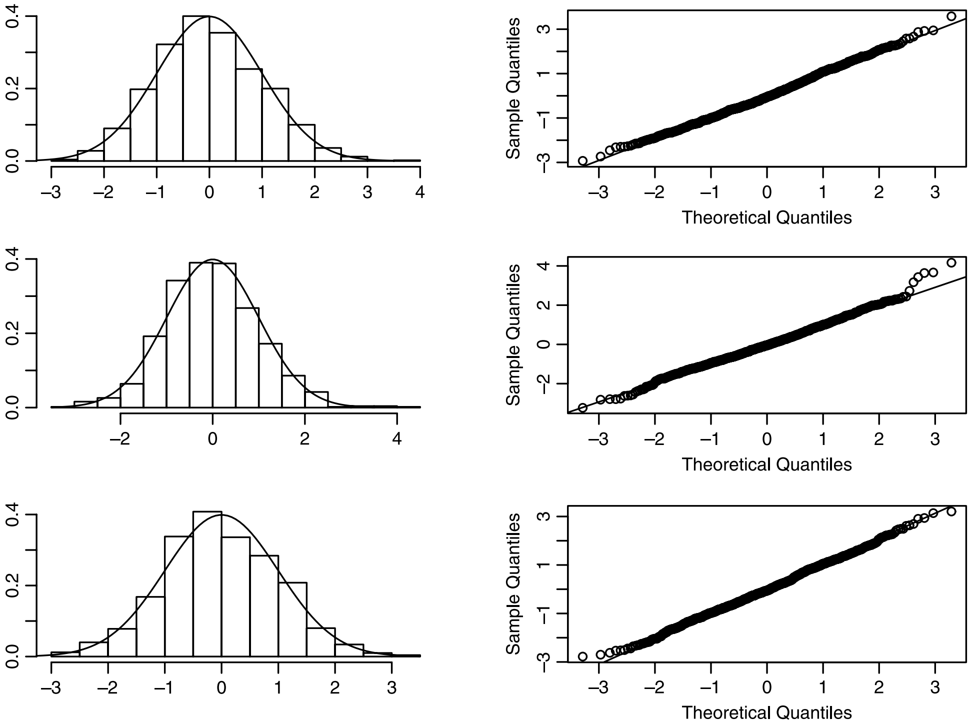

Fig. 2. From top to bottom: histograms and qq-plots of the standardized sampling distribution of $\hat{\theta} = (\hat{d}, \hat{a}, \hat{b})$ for the log-linear model (2) when the true values are $(d_0, a_0, b_0) = (0.50, -0.50, 0.65)$. The results are based on 500 data points and 1000 simulations. From top to bottom: $\hat{d}$; $\hat{a}$; $\hat{b}$.

Table 2
Simulation results. The third column reports the mean of estimators obtained by the maximum likelihood method. The fourth column shows the sample standard deviation of the estimators obtained by simulation. The other three columns report sample skewness, sample kurtosis and the p-value of a Kolmogorov–Smirnov test statistic (for testing against the standard normal distribution) for the standardized MLE obtained by the simulation. Results are based on 1000 simulations.

$d = 0.5, a = -0.50, b = 0.65$						
Parameters	Sample size	MLE	Standard error	Skewness	Kurtosis	p-value
d_0	200	0.501	0.187	0.269	3.226	0.443
a_0		−0.505	0.130	0.451	3.695	0.023
b_0		0.651	0.104	0.064	3.033	0.883
d_0	500	0.498	0.114	0.187	2.842	0.224
a_0		−0.497	0.081	0.208	3.502	0.578
b_0		0.649	0.063	0.087	2.942	0.556
d_0	1000	0.501	0.079	0.077	2.898	0.728
a_0		−0.500	0.055	0.155	3.254	0.936
b_0		0.649	0.045	0.022	2.819	0.477
$d = 0.5, a = -0.50, b = -0.35$						
d_0	200	0.488	0.113	−0.458	3.902	0.078
a_0		−0.375	0.303	1.655	6.775	0.000
b_0		−0.370	0.123	−0.072	3.304	0.982
d_0	500	0.492	0.066	−0.019	2.927	0.957
a_0		−0.469	0.149	1.057	6.340	0.000
b_0		−0.353	0.075	−0.112	2.843	0.674
d_0	1000	0.499	0.046	−0.109	2.851	0.806
a_0		−0.485	0.102	0.494	3.961	0.295
b_0		−0.353	0.054	−0.082	2.871	0.697

For the analysis of all time series we fit both the linear model (1) and the log-linear model (2). Set $\lambda_0 = 0$ and $\partial\lambda_0/\partial\boldsymbol{\theta} = 0$ for initializing the recursions regarding the linear model. For the log-linear model, set $\nu_0 = 1$ and $\partial\nu_0/\partial\boldsymbol{\theta} = \mathbf{0}$. Table 3 lists the results of the analysis. The numbers in parentheses, next to the estimators, correspond to their standard errors. These quantities are computed by using the robust sandwich matrix $H_n(\hat{\theta})G_n^{-1}(\hat{\theta})H_n(\hat{\theta})$, where $G_n(\hat{\theta})$ has been defined in Theorem 3.1 and $H_n(\theta)$ is given by (9). To examine the adequacy of the fit, consider the Pearson residuals given by $e_t = (Y_t - \lambda_t)/\sqrt{\lambda_t}$. Under the correct model, the sequence e_t is a white noise sequence with constant variance; see

Table 3
Data analysis results.

Linear model				Log-linear model fit			
Transaction data—two-minute interval							
$\hat{d}$	$\hat{a}$	$\hat{b}$	MSE	$\hat{d}$	$\hat{a}$	$\hat{b}$	MSE
0.471 (0.064)	0.745 (0.012)	0.222 (0.010)	2.770	0.082 (0.012)	0.725 (0.012)	0.243 (0.010)	2.822
Transaction data—five-minute interval							
$\hat{d}$	$\hat{a}$	$\hat{b}$	MSE	$\hat{d}$	$\hat{a}$	$\hat{b}$	MSE
2.495 (0.301)	0.527 (0.020)	0.403 (0.016)	4.106	0.331 (0.036)	0.437 (0.021)	0.472 (0.016)	4.196

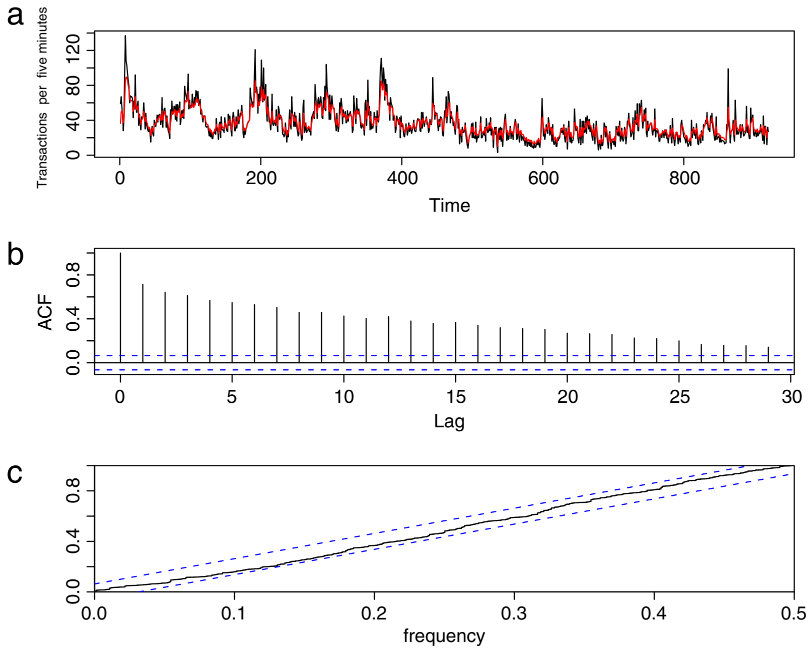

Fig. 3. (a) Transaction data aggregated over five-minute intervals. The red line corresponds to the prediction $\lambda_t(\hat{\boldsymbol{\theta}})$ obtained by fitting model (2). (b) Autocorrelation function of the data. (c) Cumulative periodogram plot of the Pearson residuals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

[15, Sec. 1.6.3]. To estimate the Pearson residuals, substitute λ_t by $\lambda_t(\hat{\boldsymbol{\theta}})$. Comparison among the models is implemented by calculating the mean square error of the Pearson residuals which is equal to $\sum_{t=1}^{N} e_t^2/(N-p)$. Here p denotes the number of estimated parameters; see [15, Sec. 1.8] for more. In addition, we plot the cumulative periodogram plot of the Pearson residuals to check for any departures from white noise; see Fig. 3(c) for the log-linear model fit.

In the two cases considered the results from the data analysis are comparable in the sense that the MSE yields similar conclusions from the two models. In fact, predictions obtained from the two models were very close—results are available by the authors. The intuitive explanation is that both models explain adequately the strong correlation, but in different scales. We anticipate that the log-linear model specification (2) will be superior to the linear model specification (1) when there is strong negative correlation in the data. However, our experience shows that such data sets are not readily available. Therefore the data analysis results obtained from the two models will be comparable because of the positive correlation among dependent counts.

Furthermore, both models yield white noise residuals. Note that the sum of estimated coefficients is close to 1 for both linear and log-linear models, regardless of the chosen time interval. This corresponds to an often observed phenomenon for GARCH(1, 1) models. To examine whether inclusion of a feedback process improves the fit of the log-linear model, we fit model (2) with $a = 0$. Then, we compare the MSE of the Pearson residuals obtained by (2) to the MSE obtained by the model without the feedback mechanism. It turns out that when $a = 0$, then the MSE is equal to 3.489 (for the two-minute interval) and 4.664 (for the five-minute interval). Hence, there is a substantial reduction in the MSE when the feedback process is introduced in the model.

Finally, to examine the sensitivity of the results as a function of the log term in model (2), we fit the following series of models to the log-mean processes:

$$\nu_t = d + a\nu_{t-1} + b\log(Y_{t-1} + \upsilon),$$

for both time series, where υ is a constant which takes values from 1 to 10 with step equal to 0.5. Initial values are set as before, that is $\nu_0 = 1$ and $\partial\nu_0/\partial\boldsymbol{\theta} = \mathbf{0}$. We calculate the MSE of the Pearson residuals for all different model specifications obtained by varying the constant υ. For the two-minute interval data, the minimum value of the MSE is equal to 2.798 (obtained at $\upsilon = 10$) while the maximum value of the MSE is 2.822 (obtained at $\upsilon = 1$). We note that the difference between these two values is negligible. A similar phenomenon occurs for the five-minute interval data. In this case, the minimum value of the MSE is equal to 4.140 (again obtained at $\upsilon = 10$) while the maximum value of the MSE is 4.196 (obtained at $\upsilon = 1$). As before, the range of all MSE values is small. Note that the sample variance for both MSE values obtained is almost zero. In conclusion, we see that the choice of $\log(Y_{t-1} + 1)$ does not affect the results of the analysis greatly, at least for these two time series.

4.3. Inclusion of covariates

As mentioned in the Introduction, one of the advantages of the model treated in this paper as compared to the model treated in [10,11] is that it is easier to introduce time dependent covariates. To be more specific, suppose that $\{X_t\}$ is some covariate time series. Then enlarging the σ-field to $\mathcal{F}_t^{Y,X,\nu} = \sigma(Y_s, X_s, \nu_0, s \le t)$ we obtain the model

$$Y_t \mid \mathcal{F}_{t-1}^{Y,X,\nu} \sim \text{Poisson}(\lambda_t), \qquad \nu_t = d + a\nu_{t-1} + b\log(Y_{t-1} + 1) + cX_t, \quad t \ge 1, \tag{11}$$

where c is a real-valued parameter. Note that a model like the above cannot be cast within the framework developed by Fokianos et al. [10,11], unless $cX_t > 0$.

If $\{X_t\}$ is itself a Markov chain, then we can construct a two-dimensional Markov chain $\{\nu_t, X_{t+1}\}$ and a corresponding three-dimensional chain with $\{Y_t\}$ included. If the transition mechanism of $\{X_t\}$ does not depend on $\{\nu_t, Y_t\}$, it is simple to find conditions for geometric ergodicity. Due to the triangular structure when $\{X_t\}$ is exogenous, separate conditions for $\{X_t\}$ are found from the transition mechanism for $\{X_t\}$, whereas the conditions for $\{\nu_t, Y_t\}$ are exactly as before. Note that if $\{X_t\}$ is a continuous-valued process it is trivial to obtain ϕ-irreducibility for $\{\nu_t\}$. We do not have to introduce a perturbed model, since in a sense $\{X_t\}$ plays the role of $\{\varepsilon_{t,m}\}$ in (5). This leads to simplification in the likelihood theory. If however $\{X_t\}$ is an integer-valued process, the perturbation argument is again needed, at least in the approach that we are using.

Inference for model (11) can proceed by means of partial likelihood theory; see [4,27] and [15, Ch. 1]. The asymptotic theory can be developed as in Section 3. For instance Eq. (7) will remain intact while recursions (8) are augmented by $\partial\nu_t/\partial c = X_t + a\partial\nu_{t-1}/\partial c$. This discussion shows that Assumption A of [15, pp. 16] can be weakened considerably in the context of modeling count time series. As a final remark, a more general model would allow $\{X_t\}$ to depend on $\{Y_t, \nu_t\}$. This is more difficult problem and will be treated in a separate publication.

We illustrate the use of covariates for the transaction data, where we have available information on the price of the stock per minute. To gain some insight into a possible relationship between the number of transactions and volatility we use the log-linear model (11) with the estimated volatility of weighted prices. More formally, suppose that P_t denotes the price per minute at time t. Then we consider the weighted averages $P_{kt} = \sum_{i=1}^{k} Y_{k(t-1)+i}P_{k(t-1)+i}/\sum_{i=1}^{k} Y_{k(t-1)+i}$, for $k = 2$ (two-minute interval) and $k = 5$ (five-minute interval) by recalling that Y_t denotes the number of transactions per minute. This is a reasonable measure for the price of the stock for each time interval because it corresponds to a price average weighted by the number of trades. If there are no transactions, then we set P_{kt} equal to the mean price of the stock over the indicated time interval.

Assume that r_{kt} stands for the log-returns of the sequence P_{kt}, that is $r_{kt} = \log(P_{kt}/P_{k(t-1)})$. We examine whether the volatility of log-returns has an impact on the number of transactions. Towards this goal we fit a GARCH(1, 1) model (see [1]) to the series r_{kt}, $k = 2, 5$. Therefore consider

$$r_{kt} = \sigma_{kt}\epsilon_t, \qquad \sigma_{kt}^2 = \omega_k + \alpha_{k1}r_{k(t-1)}^2 + \beta_{k1}\sigma_{k(t-1)}^2,$$

where ϵ_t is a sequence of independent standard normal random variables. It turns out that for the two-minute interval data, $\hat{\omega}_2 \sim 0$, $\hat{\alpha}_{21} = 0.039$ (0.008) and $\hat{\beta}_{21} = 0.922$ (0.013), where the numbers in parentheses are the standard errors of the estimators. Similarly, for the five-minute interval, we obtain that $\hat{\omega}_5 = 0.024$ (0.006), $\hat{\alpha}_{51} = 0.229$ (0.048) and $\hat{\beta}_{51} = 0.261$ (0.146). Note that the GARCH effect appears to be stronger in series r_{2t}. For each choice of time interval, the volatility is estimated by $\hat{\sigma}_{kt}^2 = \hat{\omega}_k + \hat{\alpha}_{k1}r_{k(t-1)}^2 + \hat{\beta}_{k1}\sigma_{k(t-1)}^2$. To link the volatility of the log-returns to the number of transactions, we employ model (11) with $X_{kt} = \log\hat{\sigma}_{kt}^2$, $k = 2, 5$ (note that X_{kt} may be negative). We choose again to work with the logarithm of volatility instead of $\hat{\sigma}_{kt}^2$ so that the scale remains intact among different predictors. Maximum likelihood estimation yields the following interesting results. For the two-minute interval aggregated data, $\hat{d} = 0.477(0.073)$, $\hat{a} = 0.713(0.013)$, $\hat{b} = 0.241(0.010)$ and $\hat{c} = 0.032(0.006)$. Note that the standard errors are calculated by using the sandwich estimator as in the analysis without covariates and the Pearson residuals based MSE is equal to 2.811. Hence the MSE decreases slightly when we include the covariate but more importantly these data suggest that the mean number of transactions increases as volatility increases because of the positive sign of $\hat{c}$. Similar findings were obtained for the five-minute aggregated data. In this case, $\hat{d} = 0.385(0.047)$, $\hat{a} = 0.463(0.018)$, $\hat{b} = 0.458(0.014)$, $\hat{c} = 0.032(0.014)$ and its MSE is equal to 4.179. Fig. 4 shows cumulative periodogram plots of the Pearson residuals for both time series. The graphs do not indicate any gross departures from the fitted model. The results from all of the analysis are consistent and show that the log-linear model (11) can provide a reasonable framework for the regression analysis of count dependent data.

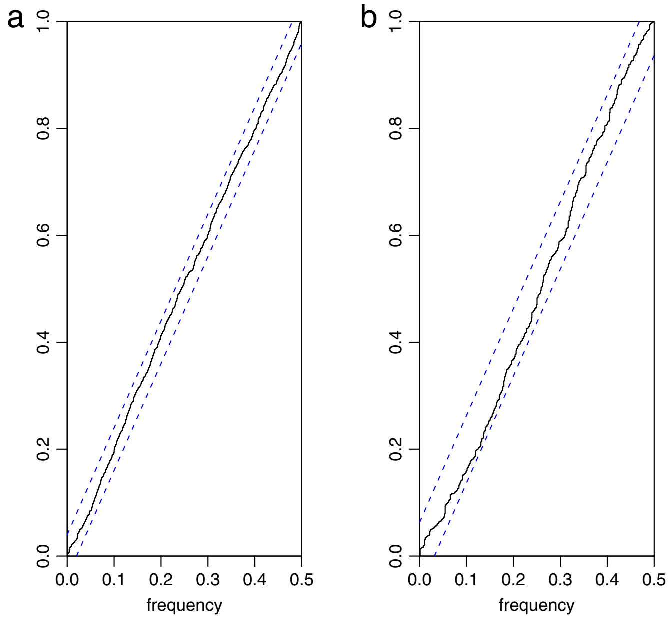

Fig. 4. Cumulative periodogram plot of the Pearson residuals for fitting model (11) to aggregated transaction data. (a) Two-minute interval. (b) Five-minute interval.

Acknowledgments

Part of this work was carried out while K. Fokianos was visiting the Institute of Mathematics, EPFL. He would like to thank all the members of the Institute for their warm hospitality and several discussions. We also appreciated the help of K. Brännäs who provided us with the transaction data and the comments of two reviewers that led to improvement of the presentation. The research was supported by Cyprus Research Foundation, grant PROSELKISI/PROEM/0308/01.

Appendix

In what follows we first prove some auxiliary lemmas and then we give the proofs of all other results. The strategy of proof is partly based upon arguments developed in [11]. In this Appendix, we concentrate on the new results and techniques needed and mostly (constrained by readability, though) refer the reader to [11] for remaining results. The first two lemmas can be combined into one, but for clarity of presentation they are given as two separate results.

Auxiliary lemmas

Lemma A.1. *For models (2) and (5),*

$$\mathrm{E}(\log(Y_t+1)|\nu_t=\nu)-\nu\to 0,$$

as $\nu\to\infty$.

Proof. We omit the upper index m for model (5) in the following proof, since it does not depend on the perturbation argument. We begin with the following identity:

$$\mathrm{E}(\log(Y_t+1)|\nu_t=\nu)=\mathrm{E}\left[\log\left(\frac{Y_t+1}{\mathrm{e}^{\nu_t}+1}(\mathrm{e}^{\nu_t}+1)\right)\Bigg|\,\nu_t=\nu\right]\sim \mathrm{E}\left[\log\left(\frac{Y_t+1}{\mathrm{e}^{\nu_t}+1}\right)\Bigg|\,\nu_t=\nu\right]+\nu,$$

as $\nu\to\infty$. We must show that the first term tends to zero as $\nu\to\infty$. Due to the Poisson property we have

$$\mathrm{E}\left(\frac{Y_t+1}{\mathrm{e}^{\nu_t}+1}\Bigg|\,\nu_t=\nu\right)=1\quad\text{and}\quad \mathrm{Var}\left[\frac{Y_t+1}{\mathrm{e}^{\nu_t}+1}\Bigg|\,\nu_t=\nu\right]\sim \mathrm{e}^{-\nu}.$$

Therefore

$$\mathrm{E}\left[\left(\frac{Y_t+1}{\mathrm{e}^{\nu_t}+1}-1\right)^2\Bigg|\,\nu_t=\nu\right]\sim \mathrm{e}^{-\nu}.$$

To ease the notation we omit the conditioning on $\nu_t = \nu$ in the rest of this proof. By the mean value theorem

$$\log\left(\frac{Y_t+1}{e^{\nu}+1}\right) = \frac{1}{\xi}\left(\frac{Y_t+1}{e^{\nu}+1} - 1\right)$$

where ξ is an intermediate point, between 1 and $(Y_t+1)/(e^{\nu}+1)$. The Schwartz inequality implies

$$E\left(\log\frac{Y_t+1}{e^{\nu}+1}\right) \le \sqrt{E\left(\frac{1}{\xi^2}\right)E\left(\frac{Y_t+1}{e^{\nu}+1} - 1\right)^2}$$

where the last term under the square root sign tends to zero with rate $e^{-\nu}$ as $\nu \to \infty$, so the proof will be completed if it can be shown that $E(1/\xi^2)$ is bounded. If $\xi \ge 1$, then $1/\xi^2 \le 1$, and

$$E\left[\frac{1}{\xi^2}1(\xi \ge 1)\right] \le 1.$$

The other alternative is $\xi < 1$, in which case

$$\frac{Y_t+1}{e^{\nu}+1} \le \xi < 1, \quad \text{or} \quad \frac{1}{\xi^2} \le \frac{(e^{\nu}+1)^2}{(Y_t+1)^2}. \tag{A.1}$$

Moreover, with $\lambda = e^{\nu}$,

$$\begin{aligned} E\left[\frac{1}{(Y_t+1)^2}1(\xi < 1)\right] &\le E\left(\frac{1}{(Y_t+1)^2}\right) = \sum_n \frac{1}{(n+1)^2}\frac{\lambda^n}{n!}e^{-\lambda} \\ &= \frac{1}{\lambda^2}\sum_n \frac{\lambda^{n+2}}{(n+1)^2 n!}e^{-\lambda} \sim \frac{1}{\lambda^2} = e^{-2\nu}. \end{aligned}$$

Hence by (A.1), $E\left(\frac{1}{\xi^2}1(\xi < 1)\right)$ is bounded, and the proof is completed. □

Lemma A.2. *For models (2) and (5),*

$$E[(\log(Y_t+1))^i|\nu_t = \nu] \sim \nu^i, \quad i = 1, 2, 3, \ldots$$

when $\nu \to \infty$.

Proof. We use the same trick as in the proof of Lemma A.1, and write

$$\begin{aligned} E[(\log(Y_t+1))^i|\nu_t = \nu] &= E\left[\left\{\log\left(\frac{Y_t+1}{e^{\nu_t}+1}\right)(e^{\nu_t}+1)\right\}^i \middle| \nu_t = \nu\right] \\ &\sim \sum_{j=1}^{i}\binom{i}{j}E\left[\left(\log\left(\frac{Y_t+1}{e^{\nu_t}+1}\right)\right)^j \middle| \nu_t = \nu\right]\nu^{i-j} + \nu^i. \end{aligned}$$

The proof will be completed if it can be shown that

$$E\left[\left\{\log\left(\frac{Y_t+1}{e^{\nu}+1}\right)\right\}^j \middle| \nu_t = \nu\right] \sim \nu^{j-1}.$$

In fact, we will show that this term tends to zero. As in the proof of Lemma A.1, we use the mean value theorem

$$\left(\log\frac{Y_t+1}{e^{\nu}+1}\right)^j = \left(\frac{1}{\xi}\right)^j\left(\frac{Y_t+1}{e^{\nu}+1} - 1\right)^j,$$

where ξ has the same interpretation as in the proof of Lemma A.1. Next, using properties of the Poisson distribution,

$$E\left[\left(\frac{Y_t+1}{e^{\nu}+1} - 1\right)^{2j} \middle| \nu_t = \nu\right] \sim \frac{(e^{\nu})^{2j-1}}{(e^{\nu}+1)^{2j}} \sim e^{-\nu} \to 0.$$

On the other hand it can be shown that $\mathrm{E}(1/\xi^{2j})$ is bounded. As in Lemma A.1, this is trivial for $\xi \geq 1$. In the case of $\xi < 1$ we need only note that

$$\mathrm{E}\left[\frac{1}{Y_t+1}\right]^{2j} = \frac{1}{\lambda^{2j}}\sum_n \frac{\lambda^{n+2j}}{(n+1)^{2j}n!}\mathrm{e}^{-\lambda} \sim \frac{1}{\lambda^{2j}} = \mathrm{e}^{-2j\nu},$$

which balances against $(\mathrm{e}^{\nu}+1)^{2j}$. Use of the Schwartz inequality then completes the proof. □

To prove Proposition 2.1 we show that $\{\nu_t^m, t \geq 0\}$ is aperiodic, ϕ-irreducible and that there exists a small set C and a test function $V(\cdot)$ which satisfies

$$\mathrm{E}[V(\nu_{t+1}^m)|\nu_t^m = \nu] \leq (1-k_1)V(\nu) + k_2 1(\nu \in C) \tag{A.2}$$

for some constants k_1, k_2 such that $0 < k_1 < 1, 0 < k_2 < \infty$. This implies that the chain $\{\nu_t^m, t \geq 0\}$ is geometrically ergodic and with a proper choice of V, the k'th moment of ν_t^m exists for an arbitrary k; see [20].

Proof of Proposition 2.1. We essentially follow the set-up of Lemma A-1 and Proposition 2.1 of [11], but unlike in that paper the parameters d, a, b may now be both positive and negative. If $b > 0$ the nonnegative term $\log(Y_{t-1}+1)$ is fed positively into the equation for ν_t, and we are in much the same situation as in [11], except that a may be positive or negative.

Proof of ϕ-irreducibility. The relevant condition is $|a| < 1$. We first consider the case of $b > 0$ and $0 < a < 1$, and let ϕ be the Lebesgue measure with support $[k, \infty)$ for some $k \geq \nu^*$. If we examine the proof of Lemma A-1 of [11], and consider a point c in $[\nu^*, \infty)$, and let $Y_1 = M$ and $N = \log(M+1)$, and replace λ by ν, then the proof can be carried through exactly as before. For $b > 0$ and $a < 0$ the same proof can be used with a replaced by $|a|$ in the majorization of $\nu_{1+j}(N) - \nu_{1+j}(N-1) = \nu_{1+j}(\log(M+1)) - \nu_{1+j}(\log(M+1)-1)$. Next we look at $b < 0$. Then the term $\log(Y_{t-1}+1)$ is fed negatively into the chain, and the ϕ-measure is chosen to be the Lebesgue measure on $(-\infty, k]$ for some $k < \nu^*$. With these changes and the above reasoning it is trivial to carry out an analogue of the proof of Lemma A-1 in [11]. This demonstrates that $\{\nu_t\}$, and $\{\nu_t^m\}$ are open set irreducible. From this, ϕ-irreducibility follows for $\{\nu_t^m\}$ (but not for $\{\nu_t\}$) using exactly the same reasoning as in the first paragraph of the proof of Proposition 2.1 of [11]. □

Proof of existence of a small set and proof of aperiodicity. These proofs can essentially be combined as in [11]. □

Proof of stability and geometric ergodicity. Stability and geometric ergodicity of the process is more difficult to prove than in [11] because of the larger flexibility in choosing a and b and because of the presence of the log term. It turns out that we need a characterization of the behavior of $E(\log(Y_t^m+1)|\nu_t = \nu)$ for ν large. Such a characterization is contained in Lemma A.1. First, we use the test function $V(x) = |x|$, and then subsequently look at the existence of arbitrary moments.

$a > 0, b > 0$: We consider first $\nu_{t-1}^m = \nu > 0$ large. Then

$$\mathrm{E}[|\nu_t^m| \mid \nu_{t-1}^m = \nu] \leq \mathrm{E}[|d + a\nu + b\log(Y_{t-1}^m+1)| \mid \nu_{t-1}^m = \nu] + \mathrm{E}(|\varepsilon_{t,m}|). \tag{A.3}$$

Using Lemma A.1 and the finiteness of d and $\mathrm{E}(|\varepsilon_{t,m}|)$,

$$E[|\nu_t^m| \mid \nu_{t-1}^m = \nu] \sim (a+b)\nu.$$

The case $\nu < 0$ with $|\nu|$ large is actually much simpler because then $\mathrm{E}[\log(Y_{t-1}^m)|\nu_{t-1}^m = \nu]$ can be made arbitrarily small by taking $|\nu|$ large enough (recall that the conditional distribution of Y_{t-1}^m is Poisson with intensity parameter $\exp(\nu)$). Using the Foster–Tweedie criterion and standard arguments this means that $\{\nu_t^m\}$ is geometrically ergodic and ϕ-irreducible for $(a+b) < 1$ in this situation.

$a < 0, b > 0$: We again consider first $\nu_{t-1}^m = \nu > 0$ large. Then (A.3) is still valid. In the case $a = -b$ there is nothing to prove by Lemma A.1. By considering separately the cases $|a| > |b|$, and $|a| < |b|$ it follows that

$$\mathrm{E}[|d + a\nu + b\log(Y_{t-1}^m+1)| \mid \nu_{t-1}^m = \nu] \sim |a+b|\nu$$

and it is needed that $|a+b| < 1$ to obtain geometric ergodicity. Note that when $|a| > |b|$, $|a+b| < 1$ is implied by the requirement $|a| < 1$ for ϕ-irreducibility. For $\nu_{t-1}^m = \nu < 0$ and $|\nu|$ large, $\mathrm{E}(\log(Y_{t-1}^m)|\nu_{t-1}^m = \nu)$ can be made arbitrarily small and $|a| < 1$ secures stability. The conclusion of all this is that for $a < 0, b > 0$, the conditions $|a| < 1$ and $|a+b| < 1$ imply ϕ-irreducibility and geometric ergodicity of $\{\nu_t^m\}$ using the Foster–Tweedie criterion.

$a < 0, b < 0$: To handle this case it is convenient to use Tjøstheim [25, Lemma 3.1] which means that it is sufficient to look at the two-step chain. Note that this technique was also used in handling the ordinary threshold time series model having different signs in two regimes; see [22]. Accordingly we take the chain $\{\nu_t^m\}$ two steps forward:

$$\begin{aligned}\nu_t^m &= d + a\nu_{t-1}^m + b\log(Y_{t-1}^m+1) + \varepsilon_{t,m} = d + ad + a(a\nu_{t-2}^m + b\log(Y_{t-2}^m+1)) \\ &\quad + b\log(Y_{t-1}^m+1) + a\varepsilon_{t-1,m} + \varepsilon_{t,m}.\end{aligned}$$

As in the above case, we can omit the term $d + ad$ and the two terms involving $\varepsilon_{t,m}$ and $\varepsilon_{t-1,m}$ since they are finite and do not depend on ν. For the remaining terms we then first consider the case of $\nu_{t-2}^m = \nu > 0$ and large. For $\nu_{t-2}^m = \nu$ we have

$$\nu_{t-1}^m = d + a\nu + b\log(Y_{t-2}+1) + \varepsilon_{t-1,m}.$$

Since $|\varepsilon_{t-1,m}| \le c_m \to 0$ and $\log(Y^m_{t-2}+1) \ge 0$, ν^m_{t-1} will be large and negative. In fact, for a fixed $\eta > 0$, $P(\nu^m_{t-1} > \eta)$ can be made arbitrarily small by taking $\nu > 0$ large enough. This means, letting f be the conditional density of ν^m_{t-1} given ν^m_{t-2}, that

$$\mathrm{E}[\log(Y^m_{t-1}+1)|\nu^m_{t-2}=\nu] = \int \mathrm{E}\{\log(Y^m_{t-1}+1)|\nu^m_{t-1}=u\}f_{\nu^m_{t-1}|\nu^m_{t-2}=\nu}(u)\mathrm{d}u$$

is finite and bounded as ν becomes large. Hence

$$\mathrm{E}[|\nu^m_t| \mid \nu^m_{t-2}=\nu] \sim \mathrm{E}[|a(a\nu + b\log(Y^m_{t-2}+1))| \mid \nu^m_{t-2}=\nu] \sim |a|\,|a+b|\nu.$$

For $\nu_{t-2} = \nu < 0$ and with $|\nu|$ large, $\nu_{t-1} \sim |a|\,|\nu|$, and taking the chain one more step we are back to the case where $a > 0$ and $b > 0$, and in that step we need $|a+b| < 1$. Putting all this together means that if $a < 0, b < 0$, the chain $\{\nu^m_t\}$ is geometrically ergodic and ϕ-irreducible if $|a| < 1$ and $|a|\,|a+b| < 1$.

$a > 0, b < 0$: For $\nu^m_{t-2} = \nu < 0$ with $|\nu|$ large, we reason as above in taking the chain one step forward, and use the same reasoning as for the case $a < 0, b > 0$ in the next step to obtain $\mathrm{E}(|\nu^m_t| \mid \nu^m_{t-2}=\nu) \sim |a|\,|a+b|\,|\nu|$. For $\nu > 0$ and large, and $|a| < |b|$, $\mathrm{E}(\nu^m_{t-1}|\nu_{t-2}=\nu) \sim (a+b)\nu < 0$, and taking the chain one step further leads to the stability condition $|a|\,|a+b| < 1$. For $a = -b$ and $|a| < 1$ there is nothing to prove. Finally, for $|a| > |b|$, $\mathrm{E}(\nu^m_{t-1}|\nu^m_{t-2}=\nu) \sim (a+b)\nu > 0$, and using arguments similar to those used above it can be shown that for $|a| > |b|$, $|a+b| < 1$ is implied by $|a| < 1$. Summing up for $a > 0, b < 0$, the chain $\{\nu_t\}$ is ϕ-irreducible and geometrically ergodic for $|a| < 1$ and $|a|\,|a+b| < 1$. □

Proof of existence of moments. Finally we address the problem of the existence of moments. We consider the test function $V(x) = 1 + x^{2k}$ where k is an integer, and as in the proof of Proposition 2.1 of [11], we need to find the one-step, and in our case also the two-step, conditional expectation of $V(\nu^m_t)$ given previous values of ν^m_t. Consider first the one-step case $(b > 0)$

$$\begin{aligned}\mathrm{E}[V(\nu^m_t)|\nu^m_{t-1}=\nu] &= 1 + \mathrm{E}[(d + a\nu + b\log(Y^m_{t-1}+1) + \varepsilon_{t,m})^{2k}|\nu^m_{t-1}=\nu] \\ &= 1 + \sum_{i=0}^{2k}\binom{2k}{i}(a\nu)^{2k-i}\mathrm{E}[(d + b\log(Y^m_{t-1}+1) + \varepsilon_{t,m})^i|\nu^m_{t-1}=\nu]. \qquad \text{(A.4)}\end{aligned}$$

Since d and $|\varepsilon_{t,m}| \le c_m$ are independent of ν, they can be neglected as ν becomes large, and we are left with the evaluation of $\mathrm{E}(\log(Y^m_{t-1}+1)^i|\nu^m_{t-1}=\nu)$. Lemma A.2 and Eq. (A.4) yield

$$\mathrm{E}(V(\nu^m_t)|\nu^m_{t-1}=\nu) = \mathrm{E}(1 + (\nu^m_t)^{2k}|\nu^m_{t-1}=\nu) = 1 + (a+b)^{2k}\nu^{2k} + \sum_{j=0}^{2k-1} c_j\nu^j$$

for some constants c_j depending on a, b, d and ε. Continuing as in the proof of Proposition 2.1 of [11] one obtains inequality (A.2) for constants k_1 and k_2 with $0 < k_1 < 1$ and $0 < k_2 < \infty$. This implies that if $b > 0$ and $|a+b| < 1$, then $\{\nu^m_t\}$ is geometrically ergodic such that any moment of ν^m_t exists. For $b < 0$ this condition can be weakened to $|a|\,|a+b| < 1$. As in the above discussion of stability and ergodicity this can be proved by taking the chain two steps forward. The details are similar to those presented there and are omitted here.

Using the fact that $\mathrm{E}[\{\log(Y_{t-1}+1)\}^{2k}|\nu_{t-1}=\nu] = \nu^{2k} + \sum_{j=0}^{2k-1} c_j\nu^j$, it is seen that obvious changes in the proof of Proposition 2.2 of [11] can be used to prove that geometric ergodicity of the $\{\nu^m_t\}$ process implies geometric ergodicity of the chain $\{(Y^m_t, U_t, \nu^m_t)\}$, or more precisely that $\{(Y^m_t, \nu^m_t, U_t), t \ge 0\}$ is a $V_{(Y,U,\nu)}$-geometrically ergodic Markov chain with $V_{Y,U,\nu}(Y, U, \nu) = 1 + \log^{2k}(1+Y) + \nu^{2k} + U^{2k}$. □

Proof of Lemma 2.1.

Proof of (1). We first look at the case $a > 0, b > 0$. Then $a + b = |a+b| = |a| + |b|$. We have

$$\mathrm{E}|\nu^m_t - \nu_t| \le a\mathrm{E}|\nu^m_{t-1} - \nu_{t-1}| + b\mathrm{E}|\log(Y^m_{t-1}+1) - \log(Y_{t-1}+1)| + \mathrm{E}|\varepsilon_{t,m}|.$$

We need to evaluate the log difference term. Without loss of generality, we assume that $\lambda^m_{t-1} \ge \lambda_{t-1}$. We can use the representation (4) as in [11] but the structure of model (2) requires a new derivation. By Jensen's inequality

$$\mathrm{E}|\log(Y^m_t+1) - \log(Y_{t-1}+1)| \le \log\mathrm{E}\frac{Y^m_{t-1}+1}{Y_{t-1}+1} = \log\left(E\frac{\Delta Y^m_{t-1}}{Y_{t-1}+1} + 1\right)$$

where $\Delta Y^m_{t-1} = Y^m_{t-1} - Y_{t-1} = N(\lambda^m_{t-1}) - N(\lambda_{t-1})$. Due to the properties of the Poisson process, conditional on $\lambda^m_{t-1}, \lambda_{t-1}$, the increment ΔY^m_{t-1} is independent of Y_{t-1}, so

$$\log\mathrm{E}\left[\frac{\Delta Y^m_{t-1}}{Y_{t-1}+1} + 1|\lambda_{t-1}, \lambda^m_{t-1}\right] = \log\left\{\mathrm{E}\left[\frac{1}{Y_{t-1}+1}|\lambda_{t-1}, \lambda^m_{t-1}\right]\mathrm{E}[\Delta Y^m_{t-1}|\lambda_{t-1}, \lambda^m_{t-1}] + 1\right\}.$$

But

$$\mathrm{E}\left[\frac{1}{Y_{t-1}+1}|\lambda_{t-1},\lambda_{t-1}^m\right] = \sum_{n=0}^{\infty}\frac{1}{n+1}\frac{(\lambda_{t-1})^n}{n!}\mathrm{e}^{-\lambda_{t-1}}$$

$$= \frac{1}{\lambda_{t-1}}\sum_{n=0}^{\infty}\frac{(\lambda_{t-1})^{n+1}}{(n+1)!}\mathrm{e}^{-\lambda_{t-1}} = \frac{1}{\lambda_{t-1}}(1-\mathrm{e}^{-\lambda_{t-1}}) \le \frac{1}{\lambda_{t-1}}.$$

Moreover, $\mathrm{E}[\Delta Y_{t-1}^m|\lambda_{t-1},\lambda_{t-1}^m] = \lambda_{t-1}^m - \lambda_{t-1}$, so

$$\log\mathrm{E}\left[\left(\frac{\Delta Y_{t-1}^m}{Y_{t-1}+1}+1\right)\middle|\lambda_{t-1},\lambda_{t-1}^m\right] \le \log\left(\frac{\lambda_{t-1}^m-\lambda_{t-1}}{\lambda_{t-1}}+1\right) = \log\left(\frac{\lambda_{t-1}^m}{\lambda_{t-1}}\right) = \nu_{t-1}^m - \nu_{t-1},$$

which means (using $\mathrm{E}(\cdot) = \mathrm{E}[\mathrm{E}(\cdot)|\lambda_{t-1},\lambda_{t-1}^m]$)

$$\mathrm{E}|\nu_t^m - \nu_t| \le (a+b)\mathrm{E}|\nu_{t-1}^m - \nu_{t-1}| + \mathrm{E}|\varepsilon_{t,m}|.$$

Continuing this recursion,

$$\mathrm{E}|\nu_t^m - \nu_t| \le (a+b)^t(\nu_0^m - \nu_0) + \sum_{i=0}^{t-1}(a+b)^i\mathrm{E}|\varepsilon_{t-i,m}|$$

and because $\mathrm{E}|\varepsilon_{t,m}| < c_m$ where $c_m \to 0$, this means that if we choose $\nu_0^m = \nu_0$, $\mathrm{E}|\nu_t^m - \nu_t|$ can be made arbitrarily small. Since the rate at which $c_m \to 0$ is at our disposal, we can choose c_m such that $\sum c_m < \infty$, and we have almost sure convergence of ν_t^m to ν_t (and uniformly in t); i.e., $|\nu_t^m - \nu_t| < \delta_{1,m}$ almost surely for m large enough.

The case $a < 0, b < 0$ can be handled in the same way using the fact that for arbitrary random variables X and Y, $E|aX + bY| = E|(-a)X + (-b)Y|$.

The case where a and b have different signs is somewhat more difficult. Using the above technique it is easily seen that $|a| + |b| < 1$ is a sufficient condition. However, this can be weakened to $a^2 + b^2 < 1$ by considering second moments:

$$\mathrm{E}(\nu_t^m - \nu_t)^2 = a^2\mathrm{E}(\nu_{t-1}^m - \nu_{t-1})^2 + 2ab\mathrm{E}\left[(\nu_{t-1}^m - \nu_{t-1})\log\frac{Y_{t-1}^m+1}{Y_{t-1}+1}\right]$$

$$+ b^2\mathrm{E}\left(\log\frac{Y_{t-1}^m+1}{Y_{t-1}+1}\right)^2 + 2a\mathrm{E}[(\nu_{t-1}^m - \nu_{t-1})\varepsilon_{t,m}] + 2b\mathrm{E}\left[\log\frac{Y_{t-1}^m+1}{Y_{t-1}+1}\varepsilon_{t,m}\right] + \mathrm{E}[\varepsilon_{t,m}^2].$$

Following the above reasoning we only have to consider the first three terms on the right hand side. For the third term, because $(Y_{t-1}^m + 1)/(Y_{t-1} + 1) \ge 1$, we can use Jensen's inequality and a conditioning argument as above to obtain

$$b^2\mathrm{E}\left(\log\frac{Y_{t-1}^m+1}{Y_{t-1}+1}\right)^2 \le b^2\mathrm{E}(\nu_t^m - \nu_t)^2.$$

On the other hand, in the second term, $(\nu_{t-1}^m - \nu_{t-1})\log((Y_{t-1}^m + 1)/(Y_{t-1} + 1)) \ge 0$, which implies, since a and b have different signs, that

$$2ab\mathrm{E}\left[(\nu_{t-1}^m - \nu_{t-1})\log\frac{Y_{t-1}^m+1}{Y_{t-1}+1}\right] \le 0,$$

and this rather crude analysis implies that the first three terms can be majorized by $(a^2 + b^2)\mathrm{E}(\nu_{t-1}^m - \nu_{t-1})^2$. The proof of this case of different signs can now be completed by using arguments very similar to those above.

Proof of (2). During the proof of existence of moments we have already secured the existence of $\mathrm{E}(\nu_t^m - \nu_t)^2$. The existence of moments for ν_t can be proved as for ν_m^t since the proof does not require ϕ-irreducibility. Moreover it is easily seen that this is bounded independently of m, and the result established for part (1) and Lebesgue dominated convergence imply (2).

Proof of (3). Note that

$$\mathrm{E}|\lambda_t^m - \lambda_t| = \mathrm{E}[|\mathrm{e}^{\nu_t^m} - \mathrm{e}^{\nu_t}|] = \mathrm{E}[|\mathrm{e}^{\nu_t^m}(1 - \mathrm{e}^{\nu_t - \nu_t^m})|].$$

From the last part of the proof of (1), it follows that the last factor in the above expectation can be made arbitrarily small almost surely. Then (3) is proved by Lebesgue dominated convergence if we can show that $\mathrm{E}(\mathrm{e}^{\nu_t}) < \infty$ and $\mathrm{E}(\mathrm{e}^{\nu_t^m})$ is bounded independently of m. Note that it is sufficient to prove this under the condition of the lemma, but it will be seen that it can be established under the weaker conditions of Proposition 2.1. Following Meyn and Tweedie [20] (see also [26]), it is sufficient to establish the inequality (A.2) for $V(x) = \mathrm{e}^x$, i.e., to prove

$$\mathrm{E}(\mathrm{e}^{\nu_t^m}|\nu_{t-1}^m = \nu) \le (1 - k_1)\mathrm{e}^{\nu} + k_2 1(\nu \in C).$$

It is trivial to prove boundedness on C, and using the technique employed in proving (A.2) in Proposition 2.1, it is seen that for $|\nu|$ large $\mathrm{E}(|\nu_t^m|^k \mid \nu_{t-1}^m = \nu) \le (|a+b|)^k|\nu|^k + o(|\nu|^k)$ for $b > 0$ and with $|a+b|^k$ replaced by $(|a|\,|a+b|)^k$ when $b < 0$. It follows that for $b > 0$, $\mathrm{E}(\mathrm{e}^{\nu_t^m}|\nu_{t-1}^m = \nu) \le \mathrm{e}^{(|a+b|)\nu} = \mathrm{e}^{(|a+b|-1)\nu}\mathrm{e}^{\nu}$ independently of m, as ν gets large with the appropriate sign, and the desired result now follows. Obvious changes take care of the case $b < 0$. The proof of $\mathrm{E}(\mathrm{e}^{\nu_t}) < \infty$ is identical since ϕ-irreducibility is not required. The remaining parts of the lemma can be proved along the lines of Fokianos et al. [10,11]. □

Proof of Theorem 3.1. To prove the above theorem, we need the following series of lemmas. These results enable us to prove asymptotic normality of $\hat{\boldsymbol{\theta}}^m$ which in turn implies asymptotic normality of $\hat{\boldsymbol{\theta}}$.

Lemma A.3. *Define the matrices*

$$\boldsymbol{G}^m(\boldsymbol{\theta}) = \mathrm{E}\left(\exp(\nu_t^m(\boldsymbol{\theta}))\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t^m}{\partial \boldsymbol{\theta}}\right)'\right) \quad \textit{and} \quad \boldsymbol{G}(\boldsymbol{\theta}) = \mathrm{E}\left(\exp(\nu_t(\boldsymbol{\theta}))\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \nu_t}{\partial \boldsymbol{\theta}}\right)'\right).$$

Under the assumptions of Theorem 3.1, the above matrices evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ satisfy $\boldsymbol{G}^m \to \boldsymbol{G}$, as $m \to \infty$. In addition, $\boldsymbol{G}^m$ and $\boldsymbol{G}$ are positive definite.

Lemma A.4. *Under the assumptions of Theorem 3.1, the score functions for the perturbed and unperturbed model and evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ satisfy the following:*

1. $\frac{1}{\sqrt{n}}\boldsymbol{S}_n^m \xrightarrow{D} \boldsymbol{S}^m := \mathcal{N}(0, \boldsymbol{G}^m)$, *as $n \to \infty$ for each $m = 1, 2 \ldots$,*
2. $\boldsymbol{S}^m \xrightarrow{D} \mathcal{N}(0, \boldsymbol{G})$ *as $m \to \infty$,*
3. $\lim_{m\to\infty} \limsup_{n\to\infty} P\left(\left\|\boldsymbol{S}_n^m - \boldsymbol{S}_n\right\| > \varepsilon\sqrt{n}\right) = 0$, *for every $\varepsilon > 0$.*

Lemma A.5. *Under the assumptions of Theorem 3.1, the Hessian matrices for the perturbed and unperturbed model and evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ satisfy the following:*

1. $\frac{1}{n}\boldsymbol{H}_n^m \xrightarrow{P} \boldsymbol{G}^m$ *as $n \to \infty$ for each $m = 1, 2 \ldots$,*
2. $\lim_{m\to\infty} \limsup_{n\to\infty} \mathrm{P}\left(\left\|\boldsymbol{H}_n^m - \boldsymbol{H}_n\right\| > \varepsilon n\right) = 0$ *for every $\varepsilon > 0$.*

Lemma A.6. *With the neighborhood $O(\boldsymbol{\theta}_0)$ defined in (10), it holds under the assumptions of Theorem 3.1 that*

$$\max_{i,j,k=1,2,3} \sup_{\boldsymbol{\theta}\in O(\boldsymbol{\theta}_0)} \left|\frac{1}{n}\sum_{t=1}^{n} \frac{\partial^3 l_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}\right| \le M_n := \frac{1}{n}\sum_{t=1}^{n} m_t$$

where $\boldsymbol{\theta}_i$ for $i = 1, 2, 3$ refers to $\boldsymbol{\theta} = d$, $\boldsymbol{\theta} = a$ and $\boldsymbol{\theta} = b$, respectively. In addition,

$$m_t = C\left(\mu_{1t}^3 + \mu_{1t}\mu_{2t} + Y_t\mu_{3t} + \mu_{3t}^3\right)$$

$$\mu_{it} = \beta_M \sum_{j=1}^{t-i} k_{j,i}\alpha_M^{j-1} Y_{t-i-j}, \quad k_{j,1} = j,\ k_{j,2} = j\,(j+1)\ \textit{and}\ k_{j,3} = j\,(j+1)\,(j+2)\,,$$

where $\alpha_M = \max\{|\alpha_L|, |\alpha_U|\}$ and $\beta_M = \max\{|\beta_L|, |\beta_U|\}$. Define correspondingly M_n^m, m_t^m and μ_{it}^m in terms of Y_t^m and let $M^m = \mathrm{E}(m_t^m)$. Then

1. $M_n^m \xrightarrow{P} M^m$, *as $n \to \infty$ for each $m = 1, 2 \ldots$,*
2. $M^m \to M$, *as $m \to \infty$, where M is a finite constant,*
3. $\lim_{m\to\infty} \limsup_{n\to\infty} P\left(\left|M_n^m - M_n\right| > \varepsilon n\right) = 0$ *for every $\varepsilon > 0$.*

All these results can be proved along the lines of Fokianos et al. [11, Lemmas 3.1–3.4].

References

[1] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics 31 (1986) 307–327.
[2] P.J. Brockwell, R.A. Davis, Time Series: Data Analysis and Theory, 2nd ed., Springer, New York, 1991.
[3] B.A. Brumback, L.M. Ryan, J.D. Schwartz, L.M. Neas, P.C. Stark, H.A. Burge, Transitional regression models with application to environmental time series, Journal of the American Statistical Association 85 (2000) 16–27.
[4] D.R. Cox, Partial likelihood, Biometrika 62 (1975) 69–76.
[5] D.R. Cox, Statistical analysis of time series: some recent developments, Scandinavian Journal of Statistics 8 (1981) 93–115.
[6] Y. Cui, R. Lund, A new look at time series of counts, Biometrika 96 (2009) 781–792.
[7] R.A. Davis, W.T.M. Dunsmuir, S.B. Streett, Observation-driven models for Poisson counts, Biometrika 90 (2003) 777–790.
[8] R. Ferland, A. Latour, D. Oraichi, Integer-valued GARCH processes, Journal of Time Series Analysis 27 (2006) 923–942.
[9] K. Fokianos, B. Kedem, Partial likelihood inference for time series following generalized linear models, Journal of Time Series Analysis 25 (2004) 173–197.

[10] K. Fokianos, A. Rahbek, D. Tjøstheim, Poisson autoregression, Journal of the American Statistical Association 104 (2009) 1430–1439.
[11] K. Fokianos, A. Rahbek, D. Tjøstheim, Poisson autoregression (complete version) 2009. Available at: http://pubs.amstat.org/toc/jasa/104/488.
[12] J. Franke, Weak dependence of functional INGARCH processes, unpublished manuscript, 2010.
[13] S.T. Jensen, A. Rahbek, Asymptotic inference for nonstationary GARCH, Econometric Theory 20 (2004) 1203–1226.
[14] R.C. Jung, M. Kukuk, R. Liesenfeld, Time series of count data: modeling, estimation and diagnostics, Computational Statistics & Data Analysis 51 (2006) 2350–2364.
[15] B. Kedem, K. Fokianos, Regression Models for Time Series Analysis, Wiley, Hoboken, NJ, 2002.
[16] A. Latour, L. Truquet, An integer-valued bilinear type model, 2008. Available at: http://hal.archives-ouvertes.fr/hal-00373409/fr/.
[17] W.K. Li, Time series models based on generalized linear models: some further results, Biometrics 50 (1994) 506–511.
[18] I.L. MacDonald, W. Zucchini, Hidden Markov and Other Models for Discrete-Valued Time Series, Chapman & Hall, London, 1997.
[19] P. McCullagh, J.A. Nelder, Generalized Linear Models, 2nd ed., Chapman & Hall, London, 1989.
[20] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Springer, London, 1993.
[21] M. Neumann, Poisson count processes: ergodicity and goodness-of-fit, Bernoulli, 2011 (forthcoming).
[22] J.D. Petruccelli, S.W. Woolford, A threshold AR(1) model, Journal of Applied Probability 21 (1984) 270–286.
[23] T.H. Rydberg, N. Shephard, A modeling framework for the prices and times of trades on the New York stock exchange, in: W.J. Fitzgerald, R.L. Smith, A.T. Walden, P.C. Young (Eds.), Nonlinear and Nonstationary Signal Processing, Isaac Newton Institute and Cambridge University Press, Cambridge, 2000, pp. 217–246.
[24] S. Streett, Some observation driven models for time series of counts, Ph.D. Thesis, Colorado State University, Department of Statistics, 2000.
[25] D. Tjøstheim, Nonlinear time series and Markov chains, Advances in Applied Probability 22 (1990) 587–611.
[26] R.L. Tweedie, The existence of moments for stationary Markov chains, Journal of Applied Probability 20 (1983) 191–196.
[27] W.H. Wong, Theory of partial likelihood, Annals of Statistics 14 (1986) 88–123.
[28] S.L. Zeger, B. Qaqish, Markov regression models for time series: a quasi-likelihood approach, Biometrics 44 (1988) 1019–1031.