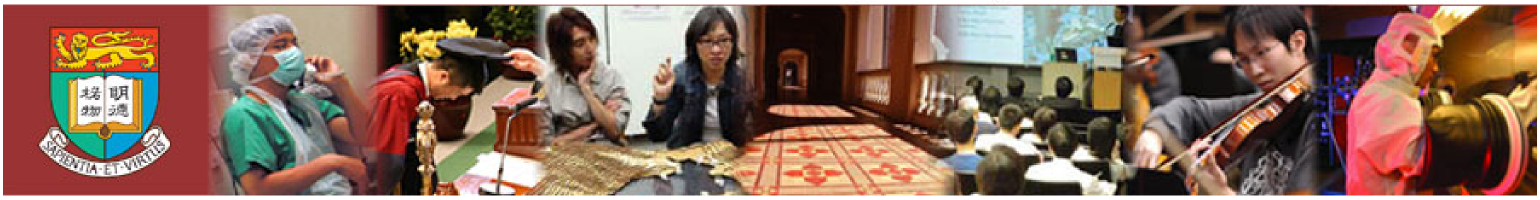

# Self-excited Threshold Poisson Autoregression

Chao Wang, Heng Liu, Jian-Feng Yao,

Richard A. Davis, Wai Keung Li [*]

## Abstract

[*]Chao Wang is Post-doctoral Fellow, Department of Statistics and Actuarial Science, the University of Hong Kong, Hong Kong (email: chaowang@connect.hku.hk); Heng Liu is Statistician, Google Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043 (email: hengliu@google.com); Jian-Feng Yao is Associate Professor, Department of Statistics and Actuarial Science, the University of Hong Kong, Hong Kong (email: jeffyao@hku.hk); Richard A. Davis is Howard Levene Professor and Chair, Department of Statistics, Columbia University, New York (email: rdavis@stat.columbia.edu); and Wai Keung Li is Professor, Department of Statistics and Actuarial Science, the University of Hong Kong, Hong Kong (email: hrntlwk@hku.hk). This research is conducted using the HKU Computer Centre research computing facilities supported partially by Hong Kong UGC Special Equipment Grant SEG HKU09. Richard A. Davis's research is supported in part by the U.S. National Science Foundation grant DMS-1107031. Wai Keung Li's research is supported partially by the HKSAR Research Grant Council General Research Fund #703711.

This paper studies theory and inference of an observation-driven model for time series of counts. It is assumed that the observations follow a Poisson distribution conditioned on an accompanying intensity process, which is equipped with a two-regime structure according to the magnitude of the lagged observations. Generalized from the Poisson autoregression, it allows more flexible, and even negative correlation, in the observations, which cannot be produced by the single-regime model. Classical Markov chain theory and Lyapunov's method are utilized to derive the conditions under which the process has a unique invariant probability measure and to show a strong law of large numbers of the intensity process. Moreover the asymptotic theory of the maximum likelihood estimates of the parameters is established. A simulation study and a real data application are considered, where the model is applied to the number of major earthquakes in the world.

**Keywords**: Integer-valued GARCH; Invariant probability measure; Self-excited threshold process; Strong law of large numbers; Time series of counts.

# 1 Introduction

There has been increasing interest in developing models for time series of counts because of their wide range of applications, including epidemiology, finance, disease modeling and environmental science. The majority of these models assume that the observations follow a Poisson distribution conditioned on an accompanying intensity process that drives the dynamics of the model, see Davis et al. (2003), Ferland et al. (2006), Fokianos et al. (2009), Fokianos and Tjøstheim (2011), Davis and Liu (2012) and Doukhan et al. (2012). According to whether the evolution of the intensity process depends on the observations or solely on an external process, Cox (1981) classified the models into observation-driven and parameter-driven. Compared to parameter-driven models, an observation-driven model usually enjoys a considerably easier and more straightforward estimation procedure, however, it is difficult to establish stability properties, including stationarity and mixing conditions of the model. This paper formulates and investigates a self-excited threshold Poisson autoregression process, which belongs to the class of observation-driven models.

One observation-driven model, the Poisson autoregression, also known as the Poisson integer-valued GARCH (INGARCH), has already received considerable study in the literature, see for example, Ferland et al. (2006), Fokianos et al. (2009), Neumann (2011), Doukhan et al. (2012), Davis and Liu (2012), and Fokianos and Tjøstheim (2012). For this model, it is assumed that the observations $\{Y_t\}$ given the intensity process $\{\lambda_t\}$ follow Poisson distribution, where $\lambda_t$ follows the GARCH-like recursions $\lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1}$.

The name GARCH associated with this model comes from Bollerslev (1986) as the Poisson mean coincides with its variance, and is known for its capability of capturing positive temporal dependence in the observations and it is relatively easy to fit via maximum likelihood. Fokianos et al. (2009) studied the model and established the asymptotic theory of the parameter estimates by introducing a small perturbation. Neumann (2011) considered some contracting dynamics of $\lambda_t$ and derived mixing condition of the count process. Davis and Liu (2012) generalized the conditional distribution of $\{Y_t\}$ to a one-parameter exponential family and took advantage of the theory for iterated random functions (Diaconis and Freedman, 1999; Wu and Shao, 2004) to establish stationarity and absolute regularity of the process, as well as the asymptotic distribution of the parameter estimates. Doukhan et al. (2012) showed similar results by utilizing the concept of $\tau$-weak dependence. More recently, Blasques et al. (2012) considered a class of generalized autoregressive score processes which includes Poisson autoregression as a special case and used the Dudley entropy integral to obtain a wider non-degenerate parameter region that guarantees the stationarity and ergodicity of the processes.

Despite many advantages that the Poisson autoregression model enjoys, it is incapable of modeling negative serial dependence in the observations. This can be seen through the fact that $\{Y_t\}$ can be represented as an ARMA$(1, 1)$ process with a sequence of martingale differences as innovations and with a positive autoregressive coefficient (see e.g., Davis and Liu (2012)). This concern motivated Fokianos and Tjøstheim (2011) in part to study the so-called log-linear Poisson autoregression. Our paper proposes a self-excited threshold integer-valued Poisson autoregression model (SETPAR), which allows for a more general modeling framework for the intensity process, including the possibility of negative serial dependence in the data. The model assumes a two-regime structure of the conditional mean process $\{\lambda_t\}$ according to the magnitude of the lagged observations. Such an extension to a model with threshold has its own merits, on account of the successful modeling strategy of a self-excited threshold autoregressive moving average process introduced by Tong (1990).

Some studies have been directed to this model from different perspectives. Woodard et al. (2011) discussed a large class of the so-called "generalized autoregressive moving average models" which includes a similar threshold model. The model was also found in another general study of observation-driven time series models by Douc et al. (2013). Despite several similar results found in their papers and ours, we adopt a different methodology, which is well suited to these types of models. The difficulty with the theory is that the Markov kernel associated with the model lacks proper continuity. Woodard et al. (2011) adopted the existing approach of Fokianos et al. (2009) which is based on a smoothed approximation of the Markov chain by adding an asymptotically vanishing noise. Douc et al. (2013) considered the model directly and applied a coupling construction to prove the uniqueness of the stationary distribution with the same conditions on model coefficients for the ergodicity as ours (compare their Proposition 14 and our Theorem 2.3). We studied the model directly using a different concept of e-chain (see Chapter 6, Meyn and Tweedie (1993)), which has an asymptotic continuity property that guarantees the uniqueness of a stationary distribution with mild additional conditions. Regarding the coverage of the approaches, the coupling argument applies to the log-linear Poisson autoregressions (Fokianos and Tjøstheim, 2011; Douc et al., 2013) as well. This is however not surprising since the Markov chains in a log-linear Poisson autoregressions and SETPAR model are very similar and our approach through e-chains can also be used for a log-linear Poisson autoregression as well. In addition, we are able to establish consistency and asymptotic normality of the maximum likelihood estimates directly based on our discussion of the stability property of the model under mild conditions on the parameters.

The organization of the paper is as follows. Section 2 formulates the model and establishes its stability properties. Likelihood inference and asymptotic theory of the estimates are investigated in Section 3. Some numerical results, including a simulation study and a real data example are given in Section 4. The model is applied to the counts of major earthquakes in the world, and some diagnostic tools for assessing and comparing model performance are also given in this section. Section 5 discusses some problems which are worth further study and concludes the paper. Proofs of the key results in Sections 2 and Section 3 are deferred to the Appendix.

## 2 The model and its properties

For ease of discussion, only the first order self-excited threshold Poisson autoregression is investigated in this paper. However, the generalization to higher order model with multiple thresholds is also possible using similarly stylized arguments.

**Definition 2.1.** *A sequence of random observations $\{Y_t, t \in \mathbb{Z}\}$ is said to follow the self-excited threshold Poisson autoregression (SETPAR) model, if*

$$\mathcal{L}(Y_t \mid \mathcal{F}_{t-1}) = \mathit{Poisson}(\lambda_t), \tag{1}$$

*where $\mathcal{F}_t = \sigma\{Y_s,\ s \leq t\}$, and*

$$\lambda_t = \begin{cases} d_1 + a_1\lambda_{t-1} + b_1 Y_{t-1}, & Y_{t-1} \leq r, \\ \\ d_2 + a_2\lambda_{t-1} + b_2 Y_{t-1}, & Y_{t-1} > r, \end{cases} \tag{2}$$

*with $d_i > 0, a_i > 0, b_i > 0,\ i = 1, 2,$ and $r \in \mathbb{N}$.*

Let $\theta^{(i)} = (d_i, a_i, b_i)^\intercal$ $(i = 1, 2)$ be the regime-specific parameter vector. It is reasonable to assume $\theta^{(1)} \neq \theta^{(2)}$, since otherwise, the model is reduced to the ordinary Poisson autoregression. The intercept parameter $d_i$ is restricted to be positive to avoid a Poisson distribution with zero mean.

The dynamics of the process is governed by a two-regime scheme. In the following context, if $Y_{t-1} \leq r$ then we say $Y_t$ lies in the lower regime, denoted by $Y_t \in R_1$, where $R_1 = \{0, \ldots, r\}$; otherwise, $Y_t$ is in the upper regime, denoted by $Y_t \in R_2$, $R_2 = \mathbb{N} - R_1$.

Let $\{N_t(\cdot), t \in \mathbb{Z}\}$ be a sequence of independent Poisson processes with unit intensity. As suggested by Fokianos et al. (2009), it is sometimes convenient to treat $Y_t$ in Eq (1) as the sampling value of $N_t$ at time $\lambda_t$, i.e.,

$$Y_t = N_t(\lambda_t), \tag{3}$$

where $\lambda_t$ is the same as in Eq (2).

Although the process $\{\lambda_t\}$ as well as the joint one $\{(\lambda_t, Y_t)\}$ is a Markov chain, it is difficult to investigate the properties of these processes, mainly due to the fact that the real-valued intensity process $\lambda_t$ is a function of the real-valued $\lambda_{t-1}$ and the discrete-valued innovations $Y_{t-1}$ (see also Fokianos et al. (2009), Woodard et al. (2011)). In particular, it is easy to show that $\{\lambda_t\}$ is not a strong Feller chain even for the Poisson autoregression model without a threshold, which implies that one needs to apply more nonstandard Markov chain theory, such as Lyapunov's method and e-chains, in order to establish stability properties. Due to the importance of the concept of stability, its definition by Duflo (1997) is given below. Readers are referred to Sections 6.1-6.2 in Duflo (1997) and Section 6.4 in Meyn and Tweedie (1993) for other corresponding definitions and relevant theory of Lyapunov's method and e-chains.

**Definition 2.2.** *(Definition 6.1.1, Definition 6.1.4, Duflo (1997)) Suppose that a random sequence $\{X_n\}$ is defined on a metric space $E$ together with its Borel $\sigma$-field. $\{X_n\}$ is said to be a stable model if there exists a probability distribution $\mu$ on $E$ such that, for almost all $\omega$, the sequence of empirical distributions*

$$\Lambda_n(\omega, \cdot) = \frac{1}{n+1} \sum_{t=0}^{n} 1\{X_t(\omega) \in \cdot\}$$

*converges weakly to $\mu$. The distribution $\mu$ is the stationary distribution for the model.*

*A Markov chain is said to be stable if its state space is a metric space, and for any initial distribution $\nu$, the induced random sequence is stable with a stationary distribution independent of $\nu$.*

We begin with the following theorem establishing the stability of $\{\lambda_t\}$.

**Theorem 2.3.** *Consider the model in Definition 2.1. Assume $a_1 < 1$ and $a_2 + b_2 < 1$ , then*

1. *The Markov chain $\{\lambda_t\}$ is stable and possesses a unique invariant probability measure $\mu$, which has moments of all orders.*

2. *For any $\mu$-a.s. continuous function $\phi$ satisfying*

$$|\phi(\lambda)| \leq c(1+\lambda^k),$$

*for some power $k \geq 0$ and constant $c$, it holds that*

$$\frac{1}{n}\left[\phi(\lambda_1)+\cdots+\phi(\lambda_n)\right] \to \mu(\phi), a.s.$$

*for any initial value $\lambda_0$.*

The properties of the observed process $\{Y_t\}$ can be deduced from the properties of $\{\lambda_t\}$, as stated in the following corollary.

**Corollary 2.4.** *Suppose the assumptions of Theorem 2.3 hold, then the joint process $\{(\lambda_t, Y_t)\}$ is stable and $\{Y_t\}$ has finite moments of all orders.*

Similar to Theorem 2.3, the stability of the joint process ensures the law of large numbers holds for polynomial functions of $(\lambda_t, Y_t)$, which serves an important role in establishing the asymptotic theory of the estimators for the parameters in next section.

As is claimed that this model can produce negative autocorrelation, we conclude this section by some remarks on the autocorrelation function of this model. It turns out that an explicit formula of its autocorrelation function is very difficult to obtain, and to our best knowledge, no such result exists for time series models with thresholds. Based on the stability of the model, the claim can be verified by Monte Carlo simulations, since the sample autocorrelation is a consistent estimator for the theoretical autocorrelation. As to the theoretical property of the autocorrelation function, it can be proved that when $b_1$ is large enough, $\mathrm{E}\left(\lambda_t|\lambda_{t-1}\right)$ is a decreasing function of $\lambda_{t-1}$. Thus, it is likely that $\lambda_t$ and $\lambda_{t-1}$ will vary in opposite directions with high probability and the pair $(Y_t, Y_{t-1})$ will display a negative correlation as $Y_t = N_t(\lambda_t)$ and $Y_{t-1} = N_{t-1}(\lambda_{t-1})$.

# 3 Parameter estimation by maximum likelihood

Suppose we have a series of observations $\{Y_t\}_{t=1}^n$ generated from the self-excited threshold Poisson autoregression model and we want to estimate the parameters. Feasible approaches include the least squares estimator and the maximum likelihood estimator. Since the likelihood function for given observations $\{Y_t\}_{t=1}^n$ can be easily calculated with an initial value of $\lambda_1$ and the maximum likelihood estimator is likely to be more efficient than the least square estimator, we only discuss the maximum likelihood estimator here.

Recall that $\theta^{(i)} = (d_i, a_i, b_i)^\intercal$ is the parameter vector for the $i^{th}$ regime, $i = 1, 2$. Then $\theta = (r, \theta^{(1)\intercal}, \theta^{(2)\intercal})^\intercal$ denotes the vector of all parameters. Let

$\theta_0$ be the true parameter vector. Let $\lambda_{t,i} = d_i + a_i\lambda_{t-1} + b_iY_{t-1}$ $(i = 1, 2)$, then $\lambda_t = \sum_i \lambda_{t,i} 1\{Y_t \in R_i\}$. Since the $\lambda_t$'s have to be calculated recursively, an initial value $\lambda_1$ is needed.

Fix an arbitrary initial value of $\lambda_1$, denoted by $\tilde{\lambda}_1$. Let $\{\tilde{\lambda}_t\}_{t=2}^n$ be the sequence calculated by the recursive equation Eq (2) with the initial value $\tilde{\lambda}_1$ and the observed data $\{Y_t\}_{t=1}^n$. Then the log-likelihood function, apart from a constant, is

$$\tilde{\ell}(\theta) = \sum_{t=1}^{n} \tilde{\ell}_t(\theta),$$

where $\tilde{\ell}_t = -\tilde{\lambda}_t + Y_t \log(\tilde{\lambda}_t)$.

The maximum likelihood estimator of $\theta$ is

$$\hat{\theta} = \arg \max_{\theta \in ([0,r_*] \cap \mathbb{N}) \times \mathcal{D}} \tilde{\ell}(\theta), \tag{4}$$

where $r_*$ is some large positive integer and $\mathcal{D}$ is some compact subset of $\mathbb{R}^6$ which will be specified later.

To study the asymptotic behaviour of the estimator, we make the following assumption about the underlying process and the parameter space.

Assumption:

(A1) The observed sequence $\{Y_t\}_{t=1}^n$ is generated from the self-excited threshold Poisson autoregression process, with true parameter $\theta_0 \in ([0, r_*] \cap \mathbb{N}) \times \mathcal{D}^o$, where $\mathcal{D}^o$ is the interior of $\mathcal{D} \subset \boldsymbol{\Theta}$, and $\boldsymbol{\Theta} = \{(d_1, a_1, b_1, d_2, a_2, b_2)^\intercal \in$

$\mathbb{R}_+^6 : \; a_1 < 1, \; b_1 < 1, \; a_2 + b_2 < 1\}$, where $\mathbb{R}_+$ is the strictly positive part of the real line.

**Remark** The assumptions are quite natural and broad. Note the restriction of the parameters in the lower regime. Although it is shown in Corollary 2.4 that the joint process $\{(\lambda_t, Y_t)\}$ is stable for any $b_1 > 0$, currently it is necessary to assume $b_1 < 1$ when proving the asymptotic properties of the maximum likelihood estimators. We conjecture that the same asymptotic properties would hold for parameters with $b_1 \geq 1$ under other assumptions but leave it for future study. Nevertheless, the restricted parameter space still contains some explosive lower regime in the sense that $a_1 + b_1 > 1$.

Bearing in mind that the calculation of the log-likelihood $\tilde{\ell}(\theta)$ is based on an initial value of $\lambda_1$, in order to establish the asymptotic properties of $\hat{\theta}$, we need to show that the effect of selecting different initial value $\tilde{\lambda}_1$ is asymptotically negligible.

To see this, note that the process can also be represented as a varying-coefficient Poisson autoregression model in the sense that the coefficients of the Poisson autoregression model vary with the past observation. Specifically, for a given parameter vector $\theta$, let $d_t = \sum_{i=1}^{2} d_i 1\{Y_t \in R_i\}$, $a_t = \sum_{i=1}^{2} a_i 1\{Y_t \in R_i\}$ and $b_t = \sum_{i=1}^{2} b_i 1\{Y_t \in R_i\}$ $(t = 1, \ldots, n)$, assuming that no ambiguity shall be caused by the notation of $a_t$ and $b_t$ for $t = 1, 2$.

Then $\lambda_t = \lambda_t(\theta)$ satisfies the recursive equation,

$$\lambda_t = d_{t-1} + b_{t-1}Y_{t-1} + a_{t-1}\lambda_{t-1} \tag{5}$$

$$:= c_{t-1} + a_{t-1}\lambda_{t-1} \tag{6}$$

$$= \sum_{k=1}^{\infty}\prod_{j=1}^{k-1} a_{t-j}c_{t-k}. \tag{7}$$

Eq (6) defines a recursive equation of $\lambda_t$ assuming the process $\{Y_t\}$ and the vector $\theta$ is given. Let $\lambda_t = \lambda_t(\{Y_t\}, \theta)$ (with the same abbreviation) be the stationary solution as displayed in Eq (7). $\tilde{\lambda}_t$ can be regarded as a stationary approximation, which is used in practical estimation. Let $\ell_t(\theta) = -\lambda_t(\theta) + Y_t \log(\lambda_t(\theta))$ and $\ell = \ell(\theta) = \sum_{t=1}^{n} \ell_t(\theta)$ be the corresponding quantities calculated from the stationary solution.

The first major result is the strong consistency of $\hat{\theta}$ in Eq (4) under the two assumptions about the process.

**Theorem 3.1.** *Under the assumption (A1), $\hat{\theta}$ is strongly consistent, i.e., $\hat{\theta} \to \theta_0$ a.s.*

Since the threshold $r$ is integer-valued, the consistency of $\hat{r}$ implies that $\hat{r} = r$ eventually. Therefore, the efficiency of the other estimates with the threshold being estimated together is asymptotically the same as that when the threshold is known. We henceforth remove $r$ from the parameter vector $\theta$ and only consider a central limit theorem for the maximum likelihood estimator with known threshold $r$. Under this setting, $\tilde{\ell}$ is differentiable with respect to $\theta$, and the score function can be calculated using the varying-coefficient representation of $\lambda_t$ as in Eq (5).

The score function is

$$\tilde{S}_n(\theta) = \frac{\partial \tilde{\ell}(\theta)}{\partial \theta} = \sum_{t=1}^{n} (\frac{Y_t}{\tilde{\lambda}_t} - 1) \frac{\partial \tilde{\lambda}_t}{\partial \theta},$$

where

$$\frac{\partial \tilde{\lambda}_t}{\partial \theta} = \begin{pmatrix} \frac{\partial \tilde{\lambda}_t}{\partial \theta^{(1)}} \\ \\ \frac{\partial \tilde{\lambda}_t}{\partial \theta^{(2)}} \end{pmatrix}, \tag{8}$$

and

$$\frac{\partial \tilde{\lambda}_t}{\partial \theta^{(i)}} = (1, \tilde{\lambda}_{t-1}, Y_{t-1})^{\intercal} 1 \{Y_{t-1} \in R_i\} + a_{t-1} \frac{\partial \tilde{\lambda}_{t-1}}{\partial \theta^{(i)}}, \text{ for } i = 1, 2. \tag{9}$$

Let

$$G = \mathrm{E}\left[\frac{1}{\lambda_t}\left(\frac{\partial \lambda_t}{\partial \theta}\right)\left(\frac{\partial \lambda_t}{\partial \theta}\right)^{\intercal}\right],$$

then we state the asymptotic normality of the maximum likelihood estimator in the following theorem.

**Theorem 3.2.** *Under the assumption (A1) except that the threshold $r$ is known, the maximum likelihood estimator $\hat{\theta} = ((\hat{\theta}^{(1)})^{\intercal}, (\hat{\theta}^{(2)})^{\intercal})^{\intercal}$ is asymptotically normal,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, G^{-1}).$$

*Furthermore, the matrix $G$ can be estimated consistently by*

$$\widehat{G} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\tilde{\lambda}_t} \left( \frac{\partial \tilde{\lambda}_t}{\partial \theta} \right) \left( \frac{\partial \tilde{\lambda}_t}{\partial \theta} \right)^{\intercal}. \tag{10}$$

**Remark** Since $r \in \mathbb{N}$, $\tilde{\ell}$ is not differentiable with respect to the threshold variable $r$. In practice, the maximization of the log-likelihood function can be done in the following two steps.

Step (1): For each $r \in [0, r_*] \cap \mathbb{N}$, find $\theta_r^{(i)}$ such that

$$(\hat{\theta}_r^{(1)}, \hat{\theta}_r^{(2)}) = \arg \max_{(\theta^{(1)}, \theta^{(2)}) \in \mathcal{D}} \tilde{\ell}(r, \theta^{(1)}, \theta^{(2)}).$$

Step (2): The threshold is estimated by searching over all candidates

$$\hat{r} = \arg \max_{r \in [0, r_*] \cap \mathbb{N}} \tilde{\ell}(r, \hat{\theta}_r^{(1)}, \hat{\theta}_r^{(2)}),$$

and the final estimate for $\theta^{(i)}$ is $\hat{\theta}_{\hat{r}}^{(i)}$ $(i = 1, 2)$.

**Remark** Since the threshold is searched over the set of candidates $[0, r_*]$, the upper bound $r_*$ should be large enough so that the set includes the true threshold. However, since the computation time of the estimation procedure increases approximately linearly with respect to the number of candidates, $r_*$ cannot be too large when computation resource is limited. Also, when the bound is too broad, there might not be enough number of observations to ensure consistent estimation. A strategy frequently used in practice is to replace the upper bound $r_*$ as well as the lower bound 0 by some numbers determined based on the data (cf. Cheng et al. (2011)). Specifically, fix $\alpha_1 < \alpha_2 \in (0,1)$ and find the empirical $\alpha_i$-th quantile for $Y_t$, $\hat{q}_i$. Then the interval $[0, r_*]$ is replaced by $[\hat{q}_1, \hat{q}_2]$. The choice of the pair $(\alpha_1, \alpha_2)$ can be $(0.2, 0.8)$ or more conservatively $(0.1, 0.9)$.

# 4 Simulation study and real data analysis

We report the simulation study with two sets of parameters and one real data analysis in this section.

A two-step estimation procedure is applied as indicated in Section 3. First we fix $\alpha_1 = 0.2$ and $\alpha_2 = 0.8$ and find the empirical $\alpha_i$-quantile of $\{Y_i\}_{i=1}^n$, $\hat{q}_i$ $(i = 1, 2)$. Then, for a given threshold candidate, $r \in [\hat{q}_1, \hat{q}_2] \cap \mathbb{N}$, we supply the negative log-likelihood function and its gradient to E04UCF, a NAG Fortran subroutine designed to minimize a smooth function subject to constraints, to obtain the parameter estimate $\hat{\theta}_r$ for the given $r$. The final estimate is obtained by selecting $r$ and the corresponding $\hat{\theta}_r$ which minimizes the negative log-likelihood function.

## 4.1 Simulation study

Two sets of parameters are considered in our simulation. The true parameter values are listed under Table 2 and Table 3 respectively. The first parameter

| Lag | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| ACF | $-0.158$ | $-0.031$ | $0.011$ | $-0.022$ | $0.004$ |

Table 1: The autocorrelation function of a sample path simulated with the second parameter set and 10000 observations.

set has both regimes stationary, while the second one has an explosive lower regime and negative serial dependence, as illustrated in Table 1.

We are interested in checking the following points. The estimated threshold is expected to be identical to the true value when sample size is sufficiently large. The parameters for each regime are consistent and asymptotically normal, so we would like to see whether its sample mean and sample variance are close to the true ones. However, since no explicit form for the asymptotic variance is available, its inverse is estimated by $\widehat{G}$ as in Eq (10). For each set of parameters, 1000 sample paths are simulated. Then for each sample path, one estimate of $\theta$, $\hat{\theta}$, and one copy of the asymptotic covariance matrix $\widehat{G}^{-1}$ are obtained. By the asymptotic result and the law of large numbers we have $n\overline{\text{cov}}(\hat{\theta}) \approx \overline{\widehat{G}^{-1}}$, where $\overline{\widehat{G}^{-1}}$ is the sample mean of $\widehat{G}^{-1}$ over the 1000 replications. The sample covariance matrix is of course dependent on the length of sample path, however, $n\overline{\text{cov}}(\hat{\theta})$ and $\overline{\widehat{G}^{-1}}$ should be approximately equal to a constant matrix independent of $n$ provided that $n$ is sufficiently large.

The simulation results for the two sets of parameters are reported in Table 2 and Table 3 respectively. Some interesting observations can be made. In general, $\hat{r}$ converges to $r$ very fast. However the speed of this convergence seems to depend on other parameters. For the first set of parameters, even when $n$ is as large as 3000, $\hat{r}$ does not equal to $r$ in rare samples. However, $\hat{r}$ is identical to the true value when sample size is 500 for the second set of parameters, which is a moderate sample size for a threshold model.

The consistency and asymptotic variance of the other parameters are confirmed in both examples. The average estimated parameters are close to the true values and the accuracy increases as the sample size increases. However, the intercept parameters $d_i$ seem to have large variances, comparing to the other parameters. This phenomenon is also found in the Poisson autoregression model (Fokianos et al., 2009). In the first example, $n\overline{\text{cov}}(\hat{\theta})$ and $\overline{\hat{G}^{-1}}$ match each other reasonably well. Such phenomenon is not so apparent in the second example, especially for $d_i$. This might be due to the fact that the lower regime is explosive in the second example.

| Sample size | Description | $r$ | $d_1$ | $a_1$ | $b_1$ | $d_2$ | $a_2$ | $b_2$ |
|---|---|---|---|---|---|---|---|---|
| | $\theta_0$ | 7 | 0·50 | 0·70 | 0·20 | 0·30 | 0·40 | 0·50 |
| $n = 500$ | $\overline{\hat{\theta}}$ | 6·80 | 0·63 | 0·69 | 0·18 | 0·83 | 0·37 | 0·47 |
| | $n\overline{\text{cov}}(\hat{\theta})$ | 1100 | 53 | 2·34 | 1·76 | 416 | 7·69 | 6·45 |
| | $\overline{\widehat{G}^{-1}}$ | N/A | 40·8 | 2·03 | 2·32 | 444 | 6·45 | 5·60 |
| $n = 1000$ | $\overline{\hat{\theta}}$ | 7·00 | 0·56 | 0·70 | 0·19 | 0·60 | 0·38 | 0·48 |
| | $n\overline{\text{cov}}(\hat{\theta})$ | 503·5 | 34·5 | 1·85 | 2·21 | 433 | 6·84 | 6·01 |
| | $\overline{\widehat{G}^{-1}}$ | N/A | 28·9 | 1·73 | 1·79 | 405 | 5·16 | 5·46 |
| $n = 2000$ | $\overline{\hat{\theta}}$ | 7·02 | 0·53 | 0·70 | 0·20 | 0·42 | 0·39 | 0·49 |
| | $n\overline{\text{cov}}(\hat{\theta})$ | 123 | 26·2 | 1·72 | 1·90 | 288 | 4·80 | 4·76 |
| | $\overline{\widehat{G}^{-1}}$ | N/A | 25·6 | 1·62 | 1·63 | 349 | 4·78 | 5·18 |
| $n = 3000$ | $\overline{\hat{\theta}}$ | 7·00 | 0·52 | 0·70 | 0·20 | 0·37 | 0·40 | 0·50 |
| | $n\overline{\text{cov}}(\hat{\theta})$ | 5 | 26·8 | 1·76 | 1·76 | 266 | 5·33 | 4·99 |
| | $\overline{\widehat{G}^{-1}}$ | N/A | 24·5 | 1·61 | 1·61 | 332 | 4·64 | 5·05 |

Table 2: Simulation 1. The true parameters are in the row with description $\theta_0$. For each sample size, 1000 replications are simulated. Then the mean of estimates, sample size times the variance of estimates and mean of asymptotic variances (if available) are reported respectively.

| Sample size | Description | $r$ | $d_1$ | $a_1$ | $b_1$ | $d_2$ | $a_2$ | $b_2$ |
|---|---|---|---|---|---|---|---|---|
|  | $\theta_0$ | 6 | 0·50 | 0·80 | 0·70 | 0·20 | 0·20 | 0·10 |
| $n = 500$ | $\bar{\hat{\theta}}$ | 6·00 | 0·47 | 0·82 | 0·69 | 0·32 | 0·19 | 0·09 |
|  | $n\overline{\text{cov}}(\hat{\theta})$ | 0 | 28·17 | 3·36 | 2·56 | 64·96 | 1·57 | 1·74 |
|  | $\overline{\widehat{G}^{-1}}$ | N/A | 34·05 | 3·52 | 2·52 | 133·27 | 1·73 | 1·48 |
| $n = 1000$ | $\bar{\hat{\theta}}$ | 6·00 | 0·50 | 0·81 | 0·70 | 0·28 | 0·20 | 0·09 |
|  | $n\overline{\text{cov}}(\hat{\theta})$ | 0 | 30·29 | 3·35 | 2·40 | 75·55 | 1·61 | 1·27 |
|  | $\overline{\widehat{G}^{-1}}$ | N/A | 33·65 | 3·48 | 2·52 | 133·54 | 1·73 | 1·47 |
| $n = 2000$ | $\bar{\hat{\theta}}$ | 6·00 | 0·50 | 0·80 | 0·70 | 0·23 | 0·20 | 0·10 |
|  | $n\overline{\text{cov}}(\hat{\theta})$ | 0 | 29·36 | 3·28 | 2·47 | 82·68 | 1·46 | 1·21 |
|  | $\overline{\widehat{G}^{-1}}$ | N/A | 33·32 | 3·45 | 2·50 | 133·90 | 1·74 | 1·47 |
| $n = 3000$ | $\bar{\hat{\theta}}$ | 6·00 | 0·50 | 0·80 | 0·70 | 0·22 | 0·20 | 0·10 |
|  | $n\overline{\text{cov}}(\hat{\theta})$ | 0 | 32·56 | 3·64 | 2·53 | 98·93 | 1·57 | 1·43 |
|  | $\overline{\widehat{G}^{-1}}$ | N/A | 33·12 | 3·44 | 2·50 | 133·65 | 1·73 | 1·48 |

Table 3: Simulation 2. The true parameters are in the row with description $\theta_0$. For each sample size, 1000 replications are simulated. Then the mean of estimates, sample size times the variance of estimates and mean of asymptotic variances (if available) are reported respectively.

## 4.2 Analysis of annual counts of major earthquakes in the world

In this example we study the series of annual counts of major earthquakes with magnitude 7 (inclusive) or above during 1900 – 2010, which is plotted in Figure 2. The data from 1900 to 2006 can be found in page 4 of Zucchini and MacDonald (2009), and the rest is extracted from the website of U.S. Geological Survey. The sample mean and sample variance are 19·30 and 50·37 respectively, showing considerable over-dispersion. The marginal distribution of $\{Y_t\}$ in a self-excited threshold Poisson autoregression is highly expected to be non-Poissonian. It also displays strong positive serial dependence, as can be seen in Figure 1.

The series has been studied with hidden Markov models with discrete states by Zucchini and MacDonald (2009). Here we would like to compare the performances of the Poisson autoregression (PAR) versus the self-excited threshold Poisson autoregression for this data set. In order to compare the out-of-sample performances, the first 100 observations are used to estimate the parameters, while the last 11 are used to calculate the out-of-sample mean square error (MSE), serving as an assessment to model performance. The estimation results are shown in Table 4.

The self-excited threshold Poisson autoregression outperforms the ordi-

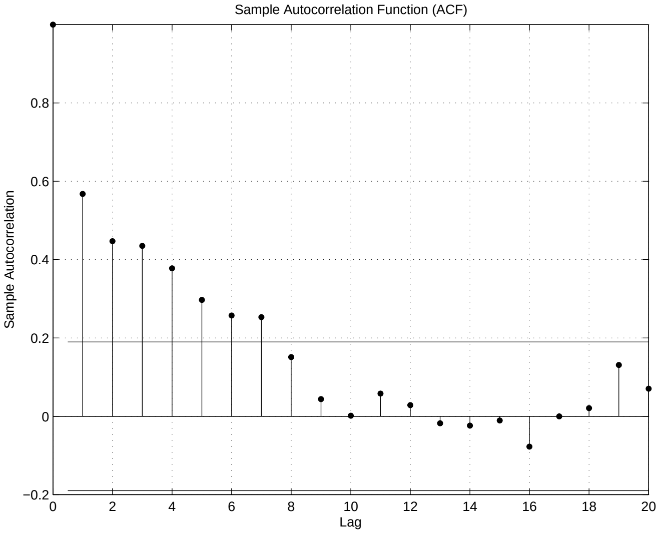

Figure 1: ACF of the earthquake data.

nary Poisson autoregression according to AIC, in-sample MSE, and out-of-sample MSE. By BIC the Poisson autoregression seems to be better, which is understandable, since BIC is very conservative when selecting models with more parameters. In the threshold case, all parameter estimates are significantly different from zero, except that $d_2$ is marginally significant and $b_2$=0·001, which in fact is the lower bound for $b_2$ in our algorithm for estimating the parameters. The same threshold model with $b_2 = 0$ is also fitted, but the result remains almost the same, as can be seen in Table 4. The basic statistics of the Pearson's residual which is defined as $(Y_t - \hat{\lambda}_t)/\sqrt{\hat{\lambda}_t}$ under the self-excited threshold Poisson autoregression model are summarized in Table 5, and its ACF is plotted in Figure 3, which shows that there is no virtually significant serial dependence in the residual sequence.

The original data and the fitted series by the two models are plotted in Figure 2. It is observed that the threshold model fits the data better when $Y_t$ is large, i.e., its improvement are mainly in the upper regime. If more data were available, a Poisson autoregression with two or more thresholds might be considered. However, insufficiency of data is very likely to result in unreliable parameter estimates, so we content ourselves with the present model.

A closer look at the fitted parameters reveals the possible different dynamics of the underlying process according to the threshold. Note that the estimated threshold is 25, which is quite large. The difference between the intercepts, $d_1$=3·27 versus $d_2$=14·33, implies that large number of major earthquakes in one year is very likely to be followed by a lot of earthquakes during the following year. Another notable feature is that $b_2 = 0$, showing that once a large number is observed, the conditional mean of the process would be stably large with less fluctuations comparing to the lower regime in which the conditional mean depends on both the latent mean process and the realized observations. For the earthquake data, this means that more earthquakes will be expected in the next few years once a large number of major earthquakes are observed in a year, as during the years 1942 – 1950 and 1968 – 1970.

| | PAR | SETPAR | SETPAR (with $b_2 = 0$) |
|---|---|---|---|
| $d_1$ | 2·96 (1·21) | 3·27 (1·36) | 3·27 (1·36) |
| $a_1$ | 0·47 (0·11) | 0·49 (0·12) | 0·49 (0·12) |
| $b_1$ | 0·39 (0·07) | 0·33 (0·10) | 0·33 (0·10) |
| $d_2$ | | 14·30 (7·45) | 14·33 (7·45) |
| $a_2$ | | 0·52 (0·20) | 0·52 (0·20) |
| $b_2$ | | 0·001 (0·26) | |
| $r$ | | 25 | 25 |
| Average log-likelihood | 39·85 | 39·89 | 39·89 |
| AIC | -7883·5 | -7885·1 | -7887·1 |
| BIC | -7875·7 | -7866·9 | -7871·5 |
| In-sample MSE | 33·12 | 30·7 | 30·7 |
| Out-of-sample MSE | 13·4 | 12·8 | 12·8 |

Table 4: Summary of model estimates. Standard errors (if available) are in parenthesis.

| Mean | Standard error | Skewness | Excess kurtosis |
|---|---|---|---|
| -0·02 | 1·219 | 0·537 | 0·429 |

Table 5: Statistics summary of the Pearson residuals of the earthquake data fitted by the self-excited threshold Poisson autoregression model.

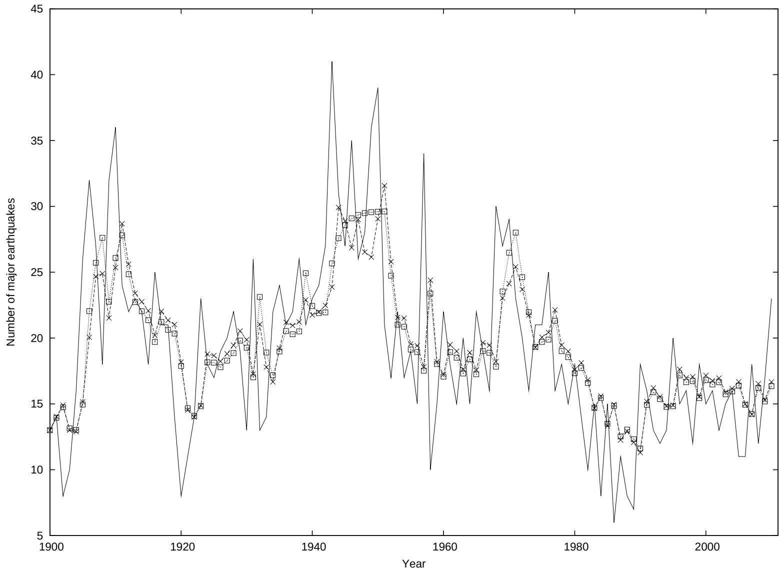

Figure 2: Plot of fitted curves of the earthquake data: The original observations are solid, the series fitted by Poisson autoregression is marked by crosses and that fitted by the self-excited threshold Poisson autoregression is marked by squares.

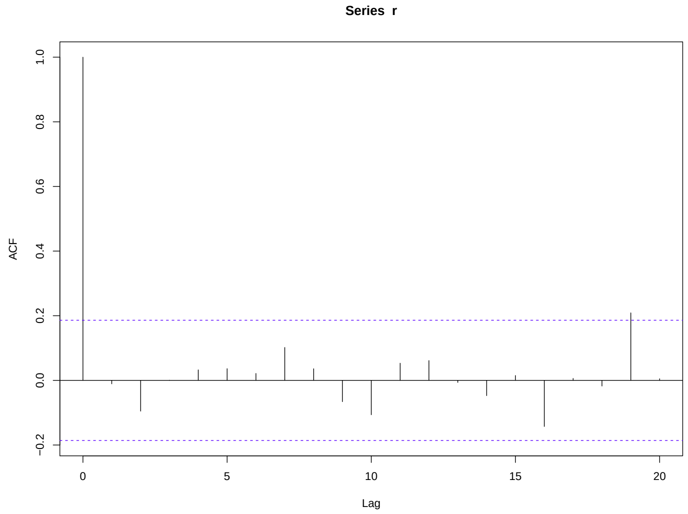

Figure 3: ACF of the Pearson residuals of the earthquake data fitted by the self-excited threshold Poisson autoregression model.

# 5 Discussion

There are some open problems deserving further investigation. The asymptotic properties of the maximum likelihood estimator derived in Theorem 3.2 might be extended to the case without the constraint that $b_1 < 1$. Another question is to test the self-excited threshold Poisson autoreregression model against the original Poisson autoregression model. Lastly, beyond the self-excited threshold Poisson model discussed in this paper, the following extension with multiple thresholds can be considered. For given integers $0 = r_0 < r_1 \cdots < r_{n-1} < r_n = \infty$, it is assumed that $\mathcal{L}(Y_t \mid \mathcal{F}_{t-1}) = \text{Poisson}(\lambda_t)$ , where

$$\lambda_t = \sum_{i=1}^{n} (d_i + a_i \lambda_{t-1} + b_i Y_{t-1}) 1\{Y_{t-1} \in [r_{i-1}, r_i)\},$$

and $d_i > 0, a_i > 0, b_i > 0$ $(i = 1, \ldots, n)$.

Results similar to Theorem 2.3,Theorem 3.1, and Theorem 3.2 can be established in a similar manner.

# 6 Appendix

In the following proofs, without explicit specification, $C$ denotes a generic positive constant, and $\rho$ a generic constant such that $\rho \in (0, 1)$. $\|X\|_p$ denotes the $L_p$-norm of a random variable $X$. The transition probability kernel of $\{\lambda_t\}$ is denoted by $\mathbf{P}$. For any function $V : \mathbb{R} \to \mathbb{R}$, let $\mathbf{P}V(\lambda) = \mathrm{E}(V(\lambda_1)|\lambda_0 = \lambda)$.

## 6.1 Proof of Theorem 2.3

*Proof.* We first prove some lemmas.

**Lemma 6.1.** *For a Poisson process $\{N(u),\ u \geq 0\}$ with unit rate,*

1. $\lim\limits_{u\to\infty} N(u)/u = 1$ *almost surely.*

2. *The family of random variables $\{(\frac{N(u)}{u})^s,\ u \geq 1\}$ is uniformly integrable for any integer $s \geq 1$.*

*Proof.* The first assertion is clearly correct for integer-valued $u$'s following the law of large numbers. For arbitrary $u$, let $\lfloor u \rfloor$ be the integer part of $u$, then $\lfloor u \rfloor \le u < \lfloor u \rfloor + 1$, and $N(\lfloor u \rfloor) \le N(u) \le N(\lfloor u \rfloor + 1)$. The conclusion follows.

For the second assertion, since $N(u)$ has a Poisson distribution with mean $u$, its $q$-th order moment is a polynomial function of $u$ of degree $q$. Therefore there exists a constant $C$ such that

$$\mathrm{E}\left(\frac{N(u)}{u}\right)^q \le C, \; u \ge 1.$$

For given order $s \ge 1$, using the bound with $q > s$ the uniformly integrability of the family $\{[N(u)/u]^s, \; u \ge 1\}$ is proved. □

**Lemma 6.2.** *For $s \ge 1$, let $V(\lambda) = \lambda^s$. Then*

$$\lim_{\lambda \to \infty} \frac{\mathbf{P}V(\lambda)}{V(\lambda)} = (a_2 + b_2)^s.$$

*Proof.* We have

$$\begin{aligned} \frac{\mathbf{P}V(\lambda)}{V(\lambda)} &= \frac{\mathrm{E}\left[V(\lambda_1) \mid \lambda_0 = \lambda\right]}{V(\lambda)} \\ &= \mathrm{E}\left[\left(\frac{d_1}{\lambda} + a_1 + b_1\frac{Y_0}{\lambda}\right)^s 1_{\{Y_0 \le r\}} + \left(\frac{d_2}{\lambda} + a_2 + b_2\frac{Y_0}{\lambda}\right)^s 1_{\{Y_0 > r\}}\right] \\ &:= \mathrm{E}[h(\lambda, \omega)] \;. \end{aligned}$$

For fixed $\omega$ and when $\lambda \to \infty$, since by Lemma 6.1, $Y_0/\lambda = N_0(\lambda)/\lambda \to 1$ a.s., $1_{\{Y_0 \le r\}} \to 0$. Therefore $h(\lambda, \omega) \to (a_2 + b_2)^s$ a.s. as $\lambda \to \infty$.

Next we check the uniform integrability condition. Using $(a+b)^s \leq 2^{s-1}(a^s+b^s)$ for $s \geq 1,\ a,b \geq 0$, it is clear that for all $\lambda \in [1,\infty)$,

$$0 \leq h(\lambda,\omega) \leq c(s)\left(1+\left(\frac{Y_0}{\lambda}\right)^s\right),$$

for some constant $c(s)$ independent of $\lambda$ (but depending on $s$ and the parameters). By Lemma 6.1, the family $\{(Y_0/\lambda)^s,\ \lambda \geq 1\}$ is uniformly integrable, so is the family $\{h(\lambda,\omega),\ \lambda \geq 1\}$. We thus obtain the announced limit. □

**Lemma 6.3.** *The Markov chain $\{\lambda_t\}$ is weakly Feller.*

*Proof.* To make the dependence on Poisson processes explicit, we write the state equation Eq (2) in the form $\lambda_t = F(\lambda_{t-1}, N_{t-1})$ with $Y_{t-1}$ replaced by $N_{t-1}(\lambda_{t-1})$, using the representation of $Y_{t-1} = N_{t-1}(\lambda_{t-1})$ in Eq (3). Let $g : \mathbb{R}_+ \to \mathbb{R}$ be any continuous and bounded function. We need to prove that $\mathbf{P}g(x) = \mathrm{E}[g(\lambda_1) \mid \lambda_0 = x]$ is continuous. Let $\varepsilon > 0$ and first choose $\eta > 0$ such that $2\|g\|_\infty(1-e^{-2\eta}) \leq \varepsilon/2$. Consider a neighbourhood $(x_0-\eta, x_0+\eta]$ of some $x_0 \in \mathbb{R}_+$. Define the event

$$A = \{\text{ the Poisson process } N_0 \text{ has no jumps in } (x_0-\eta, x_0+\eta]\ \}.$$

Clearly, $P(A) = e^{-2\eta}$. Write

$$\begin{aligned}\mathbf{P}g(x) - \mathbf{P}g(x_0) =& \mathrm{E}\left[g(F(x,N_0)) - g(F(x_0,N_0))\right] \\ =& \mathrm{E}\left[\{g(F(x,N_0)) - g(F(x_0,N_0))\}\, 1_A\right] \\ &+ \mathrm{E}\left[\{g(F(x,N_0)) - g(F(x_0,N_0))\}\, 1_{A^c}\right].\end{aligned}$$

On $A^c$, we have

$$|\mathrm{E}\left\{g(F(x,N_0)) - g(F(x_0,N_0))\right\} 1_{A^c}| \le 2\|g\|_\infty P(A^c) = 2\|g\|_\infty(1-e^{-2\eta}) \le \varepsilon/2.$$

And on the event $A$, $N_0(x) = N_0(x_0)$, for any $x \in (x_0 - \eta, x_0 + \eta]$. The mapping $x \mapsto F(x, N_0)$ is continuous, so is $x \mapsto g(F(x, N_0))1_A$ which is also bounded. Thus by Lebesgue's dominated convergence theorem,

$$\mathrm{E}\left[\left\{g(F(x,N_0)) - g(F(x_0,N_0))\right\} 1_A\right] \to 0,\ x \to x_0.$$

We can then choose $\eta_1 < \eta$ such that for $|x - x_0| < \eta_1$,

$$|\mathrm{E}\left\{g(F(x,N_0)) - g(F(x_0,N_0))\right\} 1_A| \le \varepsilon/2\ .$$

Finally for $|x - x_0| < \eta_1$, by collecting these two estimates,

$$|\mathbf{P}g(x) - \mathbf{P}g(x_0)| \le \varepsilon.$$

The proof is complete. □

**Lemma 6.4.** *The Markov chain $\{\lambda_t\}$ is an e-chain provided that $a_1 < 1$ and $a_2 + b_2 < 1$.*

*Proof.* It suffices to show that for any continuous function $f$ with compact support and $\epsilon > 0$, there exists an $\eta > 0$ such that $|\mathbf{P}^k f(x) - \mathbf{P}^k f(z)| < \epsilon$, for any $|x - z| < \eta$ and all $k \ge 1$, where $\mathbf{P}^k f(\cdot) = \mathrm{E}(f(\lambda_k) \mid \lambda_0 = \cdot)$.

Without loss of generality, assume $|f| \le 1$. Take $\epsilon'$ and $\eta$ sufficiently small such that $\epsilon' + 4\eta/(1-\bar{a}) < \epsilon$, where $\bar{a} = \max\{a_1, a_2\} < 1$, and $|f(x_1) - f(z_1)| <$

$\epsilon'$ whenever $|x_1 - z_1| < \eta$. Denote $p(\cdot \mid x)$ as the probability mass function of a Poisson distribution with intensity $x$. Then for the case when $k = 1$,

$$\begin{aligned}
& |\mathbf{P}f(x_1) - \mathbf{P}f(z_1)| \\
\leq\ & |\sum_{i=0}^{r} f(d_1 + a_1x_1 + b_1i)p(i \mid x_1) - \sum_{i=0}^{r} f(d_1 + a_1z_1 + b_1i)p(i \mid z_1)| \\
& +|\sum_{j=r+1}^{\infty} f(d_2 + a_2x_1 + b_2j)p(j \mid x_1) - \sum_{j=r+1}^{\infty} f(d_2 + a_2z_1 + b_2j)p(j \mid z_1)| \\
:=\ & I + II.
\end{aligned}$$

For $x_1 \geq z_1$,

$$\begin{aligned}
\sum_{i=0}^{\infty} |p(i \mid x_1) - p(i \mid z_1)| &= \sum_{i=0}^{\infty} |\frac{x_1^i e^{-x_1}}{i!} - \frac{z_1^i e^{-z_1}}{i!}| \\
&\leq \sum_{i=0}^{\infty} \frac{(x_1^i - z_1^i)e^{-x_1}}{i!} + \sum_{i=0}^{\infty} \frac{z_1^i(e^{-z_1} - e^{-x_1})}{i!} \\
&= 2(1 - e^{-|x_1 - z_1|}).
\end{aligned}$$

The same inequality holds for $x_1 < z_1$ by symmetry. Hence for any $x_1$ and $z_1$, we have

$$\sum_{i=0}^{\infty} |p(i \mid x_1) - p(i \mid z_1)| \leq 2(1 - e^{-|x_1 - z_1|}). \tag{11}$$

It follows that

$$\begin{aligned}
I &\leq \sum_{i=0}^{r} |f(d_1 + a_1x_1 + b_1i) - f(d_1 + a_1z_1 + b_1i)|p(i \mid x_1) \\
&\quad + \sum_{i=0}^{r} |f(d_1 + a_1z_1 + b_1i)||p(i \mid x_1) - p(i \mid z_1)| \\
&\leq \epsilon' F(r \mid x_1) + 2(1 - e^{-|x_1 - z_1|}),
\end{aligned}$$

where $F(r \mid x_1) = \sum_{i=0}^{r} p(i \mid x_1)$. The last inequality follows from Eq (11), $|f| \leq 1$, and the fact that $|(d_1+a_1x_1+b_1i)-(d_1+a_1z_1+b_1i)| = a_1|x_1-z_1| < \eta$. It follows from a similar argument that $II \leq \epsilon'(1-F(r \mid x_1))+2(1-e^{-|x_1-z_1|})$. Hence we have

$$|\mathbf{P}f(x_1) - \mathbf{P}f(z_1)| \leq \epsilon' + 4(1 - e^{-|x_1-z_1|}), \tag{12}$$

for $|x_1 - z_1| < \eta$. For the case when $k = 2$, it follows from

$$\mathrm{E}\{f(\lambda_2) \mid \lambda_0 = x\} = \mathrm{E}\{\mathrm{E}[f(\lambda_2) \mid \lambda_1] \big| \lambda_0 = x\}$$

that

$$\begin{aligned} |\mathbf{P}^2 f(x_1) - \mathbf{P}^2 f(z_1)| &= |\mathbf{P}(\mathbf{P}f)(x_1) - \mathbf{P}(\mathbf{P}f)(z_1)| \\ &\leq |\sum_{i=0}^{r} p(i \mid x_1)\mathbf{P}f(x_2^{(1)}) - \sum_{i=0}^{r} p(i \mid z_1)\mathbf{P}f(z_2^{(1)})| \\ &\quad +|\sum_{j=r+1}^{\infty} p(j \mid x_1)\mathbf{P}f(x_2^{(2)}) - \sum_{j=r+1}^{\infty} p(j \mid z_1)\mathbf{P}f(z_2^{(2)})| \\ &:= III + IV, \end{aligned}$$

where $x_2^{(1)} = d_1 + a_1x_1 + b_1i$, $x_2^{(2)} = d_2 + a_2x_1 + b_2j$, $z_2^{(1)} = d_1 + a_1z_1 + b_1i$, and $z_2^{(2)} = d_2 + a_2z_1 + b_2j$. Then

$$\begin{aligned} III &\leq \sum_{i=0}^{r} p(i \mid x_1)|\mathbf{P}f(x_2^{(1)}) - \mathbf{P}f(z_2^{(1)})| + \sum_{i=0}^{r} |\mathbf{P}f(z_2^{(1)})||p(i \mid x_1) - p(i \mid z_1)| \\ &\leq \left\{\epsilon' + 4\left(1 - e^{-|x_2^{(1)}-z_2^{(1)}|}\right)\right\} F(r \mid x_1) + 2\left(1 - e^{-|x_1-z_1|}\right), \end{aligned}$$

which follows from (11) and (12). Similarly, we have

$$IV \leq \left\{\epsilon' + 4(1 - e^{-|x_2^{(2)}-z_2^{(2)}|})\right\} (1 - F(r \mid x_1)) + 2\left(1 - e^{-|x_1-z_1|}\right).$$

Since $|x_2^{(1)} - z_2^{(1)}| = a_1|x_1 - z_1|$ and $|x_2^{(2)} - z_2^{(2)}| = a_2|x_1 - z_1|$, so by letting $\bar{a} = \max\{a_1, a_2\}$, we have

$$|\mathbf{P}^2 f(x_1) - \mathbf{P}^2 f(z_1)| \;\leq\; \epsilon' + 4\left(1 - e^{-\bar{a}|x_1 - z_1|}\right) + 4\left(1 - e^{-|x_1 - z_1|}\right).$$

Inductively, one can show that for any $k \geq 1$,

$$\begin{aligned} |\mathbf{P}^k f(x_1) - \mathbf{P}^k f(z_1)| &\leq \epsilon' + 4\sum_{s=0}^{k-1}\left(1 - e^{-\bar{a}^s|x_1 - z_1|}\right) \\ &\leq \epsilon' + 4\sum_{s=0}^{\infty}\bar{a}^s|x_1 - z_1| \\ &\leq \epsilon' + \frac{4\eta}{1 - \bar{a}} < \epsilon, \end{aligned}$$

where the second inequality holds since $1 - e^{-x} \leq x$. Hence $\{\lambda_t\}$ is an e-chain.

□

**Proof of Theorem 2.3** By Lemma 6.2, for any initial value $\lambda_0 = x$, the sequence of transition probabilities

$$\overline{\pi}_n(x, dy) = \frac{1}{n}\left\{\mathbf{P}(x, dy) + \cdots + \mathbf{P}^n(x, dy)\right\}$$

is tight (Duflo, 1997, Proposition 2.1.6). Moreover, using the weak Feller property established in Lemma 6.3, we know that the weak limit of any subsequence of $\{\overline{\pi}_n(x, dy)\}$ is an invariant probability measure of $\mathbf{P}$.

Then note that $\lambda^* = d_1/(1 - a_1)$ is a reachable state by letting $Y_1 = Y_2 = \ldots = Y_t = 0$ for large $t$. Combined with the fact that $\{\lambda_t\}$ is an e-chain, it follows that the stationary distribution is unique.

The fact that $\mu(|x|^s) < \infty$ for all $s \geq 0$ directly results from the Lyapounov property established in Lemma 6.2. The strong law of large numbers also follows from this method, see Proposition 6.2.12 and the remarks in Section 6.2.2 in Duflo (1997). The proof is complete. □

## 6.2 Proof of Corollary 2.4

*Proof.* The stability of the joint process is clear. To see $Y_t \in L_s$, for all $s > 0$, it suffices to note that $\lambda_t \in L_s$ for all $s > 0$ and the following fact

$$\mathrm{E}(Y_t)^s = \mathrm{E}[\mathrm{E}\{(Y_t)^s \mid \lambda_t\}] = (\mathrm{E}(Poly(\lambda_t, s)) < \infty,$$

where $Poly(\lambda_t, s)$ is the polynomial of $\lambda_t$ of order $s$ which represents the $s$th moment of a Poisson random variable with mean $\lambda_t$.

□

## 6.3 Proof of Theorem 3.1

*Proof.* Since the log-likelihood $\tilde{\ell}$ is calculated with a given initial value $\tilde{\lambda}_1$, we first show that the log-likelihood $\tilde{\ell}$ is asymptotically independent of $\tilde{\lambda}_1$.

Using the varying-coefficient representation in Eq (5), we have

$$\lambda_t(\lambda_1) = \sum_{k=1}^{t-2} \prod_{j=1}^{k-1} a_{t-j} c_{t-k} + \prod_{j=1}^{t-1} a_{t-j} \lambda_1,$$

which implies

$$\sup_{\theta\in\mathcal{D}}|\lambda_t(\lambda_1)-\tilde{\lambda}_t(\tilde{\lambda}_1)| = \sup_{\theta\in\mathcal{D}}|\prod_{j=1}^{t-1} a_{t-j}(\lambda_1-\tilde{\lambda}_1)| \leq K\rho^t,$$

where $\rho = \sup_{\theta\in\mathcal{D}}\max\{a_1, a_2\} < 1$ and $K = |\lambda_1-\tilde{\lambda}_1|/\rho$.

Then the difference between the log-likelihoods based on arbitrary initial value and on the stationary initial one is

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}}|\frac{1}{n}(\ell(\lambda_1)-\ell(\tilde{\lambda}_1)| =& \sup_{\theta\in\mathcal{D}}|\frac{1}{n}\sum_{t=1}^{n} Y_t(\log(\lambda_t)-\log(\tilde{\lambda}_t))-(\lambda_t-\tilde{\lambda}_t)| \\
=& \sup_{\theta\in\mathcal{D}}|\frac{1}{n}\sum_{t=1}^{n} Y_t\log(1+\frac{\lambda_t-\tilde{\lambda}_t}{\tilde{\lambda}_t})-(\lambda_t-\tilde{\lambda}_t)| \\
\leq& \sup_{\theta\in\mathcal{D}}\frac{1}{n}\sum_{t=1}^{n} Y_t|\frac{\lambda_t-\tilde{\lambda}_t}{\tilde{\lambda}_t}|+|\lambda_t-\tilde{\lambda}_t| \\
\leq& \sup_{\theta\in\mathcal{D}}\frac{1}{n}\sum_{t=1}^{n}|\lambda_t-\tilde{\lambda}_t|(\frac{Y_t}{d_0}+1) \\
\leq& \frac{1}{n}\sum_{t=1}^{n} K\rho^t(\frac{Y_t}{d_0}+1) \\
\to& 0,\ a.s.
\end{aligned}$$

where $d_0 = \inf_{\theta\in\mathcal{D}}\min\{d_1, d_2\} > 0$.

The $a.s.$ limit holds because of the Cesàro lemma and the observation that $\rho^t Y_t \to 0, a.s.$ (see also Francq and Zakoïan (2004)).

Secondly, we prove that $\mathrm{E}[\ell_t(\theta)]$ is continuous in $\theta$. Since $r$ is discrete, we need only to prove the following property. For any $\theta \in \mathcal{D}$, let $V_\eta(\theta) = B(\theta, \eta)$

be an open ball centered at $\theta$ with radius $\eta$, then

$$\mathrm{E}\left(\sup_{\tilde{\theta}\in V_\eta(\theta)} |\ell_t(\tilde{\theta}) - \ell_t(\theta)|\right) \to 0, \text{ as } \eta \to 0. \tag{13}$$

To see this, observe that

$$|\ell_t(\tilde{\theta}) - \ell_t(\theta)| \le (\frac{Y_t}{\lambda_t(\tilde{\theta})} + 1)|\lambda_t(\tilde{\theta}) - \lambda_t(\theta)|,$$

and

$$\begin{aligned}
|\lambda_t(\theta) - \lambda_t(\tilde{\theta})| =& |\sum_k \prod_{j=1}^{k-1} a_{t-j}c_{t-k} - \prod_{j=1}^{k-1} \tilde{a}_{t-j}\tilde{c}_{t-k}| \\
=& |\sum_k (\prod_{j=1}^{k-1} a_{t-j} - \prod_{j=1}^{k-1} \tilde{a}_{t-j})c_{t-k} + \prod_{j=1}^{k-1} \tilde{a}_{t-j}(c_{t-k} - \tilde{c}_{t-k})| \\
\le& C\eta \sum_k \rho^k(1 + Y_{t-k}).
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{E}\left(\sup_{\tilde{\theta}\in V_\eta(\theta)} |\ell_t(\tilde{\theta}) - \ell_t(\theta)|\right) \le& \|\frac{Y_t}{d_0} + 1\|_2 \|\lambda_t - \tilde{\lambda}_t\|_2 \\
\le& C\eta \|\frac{Y_t}{d_0} + 1\|_2 \sum_k \rho^k \|Y_t\|_2 \\
\to& 0, \text{ as } \eta \to 0.
\end{aligned}$$

Next, we check the model identifiability. By Jensen inequality, we have

$$\begin{aligned}
\mathrm{E}\left[\ell_t(\theta) - \ell_t(\theta_0)\right] =& \mathrm{E}\left[\mathrm{E}\left(\log \frac{\phi(Y_t \mid \lambda_t(\theta))}{\phi(Y_t \mid \lambda_t(\theta_0))} \mid \mathcal{F}_{t-1}\right)\right] \\
\le& \mathrm{E}\left[\log \mathrm{E}\left(\frac{\phi(Y_t \mid \lambda_t(\theta))}{\phi(Y_t \mid \lambda_t(\theta_0))} \mid \mathcal{F}_{t-1}\right)\right] \\
=& \mathrm{E}(\log(1)) = 0,
\end{aligned}$$

where $\phi(\cdot \mid y)$ denotes the Poisson distribution function with mean $y$, and the equality holds iff $\lambda_t(\theta) = \lambda_t(\theta_0)$ $a.s.$ $\mathcal{F}_{t-1}$.

Suppose that $\tilde{\theta}$ satisfies $\tilde{\lambda}_t = \lambda_t(\tilde{\theta}) = \lambda_t(\theta_0)$ $a.s.$ $\mathcal{F}_{t-1}$. Without loss of generality, assume $\tilde{r} \geq r$. For ease of notation, let $\lambda_t = \lambda_t(\theta_0)$ temporarily, then conditional on $\mathcal{F}_{t-2}$, we have $\tilde{\lambda}_{t-1} = \lambda_{t-1}$ $a.s.$, and almost surely

$$\begin{aligned}\tilde{\lambda}_t - \lambda_t =&(\tilde{d}_{t-1} + \tilde{b}_{t-1}Y_{t-1} + \tilde{a}_{t-1}\tilde{\lambda}_{t-1}) - (d_{t-1} + b_{t-1}Y_{t-1} + a_{t-1}\lambda_{t-1})\\ =&[(\tilde{d}_1 - d_1) + (\tilde{b}_1 - b_1)Y_{t-1} + (\tilde{a}_1 - a_1)\lambda_{t-1}]1\left\{Y_{t-1} \leq r\right\}\\ &+ [(\tilde{d}_1 - d_2) + (\tilde{b}_1 - b_2)Y_{t-1} + (\tilde{a}_1 - a_2)\lambda_{t-1}]1\left\{r < Y_{t-1} \leq \tilde{r}\right\}\\ &+ [(\tilde{d}_2 - d_2) + (\tilde{b}_2 - b_2)Y_{t-1} + (\tilde{a}_2 - a_2)\lambda_{t-1}]1\left\{\tilde{r} < Y_{t-1}\right\}. \quad (14)\end{aligned}$$

Note that $\mathcal{F}_{t-1} = \sigma\left\{Y_{t-1}, \mathcal{F}_{t-2}\right\}$, $Y_t \mid \lambda_t \sim \text{Poisson}(\lambda_t)$, it can be seen from Eq (14) that if $\tilde{\lambda}_t - \lambda_t = 0$ $a.s.$ $\mathcal{F}_{t-1}$, we must have $\tilde{\theta} = \theta_0$.

Now we are ready to prove the consistency. Consider an arbitrary (small) open neighbourhood of $\theta_0$, say $V$, then for any $\vartheta \in V^c \cap \mathcal{D}$, we have $\mathrm{E}[\ell_t(\vartheta)] < \mathrm{E}[\ell_t(\theta_0)]$, since $V^c \cap \mathcal{D}$ is compact and $\mathrm{E}[\ell_t(\theta)]$ is continuous in $\theta$, we have $\kappa = \mathrm{E}[\ell_t(\theta_0)] - \sup_{\theta \in V^c \cap \mathcal{D}} \mathrm{E}[\ell_t(\theta)] > 0$. And for any $\theta \in V^c \cap \mathcal{D}$, there exists $\eta_\theta > 0$ such that $\mathrm{E}[\sup_{\vartheta \in V_{\eta_\theta}(\theta)} \ell_t(\theta)] < \mathrm{E}[\ell_t(\theta)] + \frac{1}{6}\kappa$. Also by the compactness of $V^c \cap \mathcal{D}$, there exists a finite open cover of $V^c \cap \mathcal{D}$, say, $\{V_{\eta_{\theta_j}}(\theta_j), j =$

$1,\ldots,m\}$. For any $\theta \in \mathcal{D}$ and $k \gg 0$,

$$\begin{aligned}
&\overline{\lim_{n\to\infty}} \sup_{\theta^*\in V_{1/k}(\theta)\cap\boldsymbol{\Theta}} \frac{1}{n}\tilde{\ell}(\theta^*) \\
&\leq \overline{\lim_{n\to\infty}} \sup_{\theta^*\in V_{1/k}(\theta)\cap\boldsymbol{\Theta}} \frac{1}{n}\ell(\theta^*) + \overline{\lim_{n\to\infty}} \sup_{\theta^*\in V_{1/k}(\theta)\cap\boldsymbol{\Theta}} \frac{1}{n}|\ell(\theta^*) - \tilde{\ell}(\theta^*)| \\
&\leq \overline{\lim_{n\to\infty}} \frac{1}{n}\sum_{t=1}^{n} \sup_{\theta^*\in V_{1/k}(\theta)\cap\boldsymbol{\Theta}} \ell_t(\theta^*).
\end{aligned}$$

By Corollary 2.4 and as in Francq and Zakoïan (2004), we have almost surely for $n \gg 0$ and $j = 1,\ldots,m$,

$$\begin{aligned}
\sup_{\theta\in V\eta_{\theta_j}(\theta_j)} \frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(\theta) &\leq \sup_{\theta\in V\eta_{\theta_j}(\theta_j)} \frac{1}{n}\sum_{t=1}^{n}\ell_t(\theta) + \frac{1}{6}\kappa \\
&\leq \frac{1}{n}\sum_{t=1}^{n} \sup_{\theta\in V\eta_{\theta_j}(\theta_j)} \ell_t(\theta) + \frac{1}{6}\kappa \\
&\leq \mathrm{E}\left(\sup_{\theta\in V\eta_{\theta_j}(\theta_j)} \ell_t(\theta)\right) + \frac{1}{3}\kappa \\
&\leq \mathrm{E}[l_t(\theta_0)] - \frac{2}{3}\kappa.
\end{aligned}$$

And

$$\sup_{\theta\in V}\frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(\theta) \geq \frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(\theta_0) \geq \frac{1}{n}\sum_{t=1}^{n}\ell_t(\theta_0) - \frac{1}{6}\kappa \geq \mathrm{E}[\ell_t(\theta_0)] - \frac{1}{3}\kappa.$$

Therefore, for any (small) neighbourhood of $\theta_0$, $V$, for $n \gg 0$, we have almost surely

$$\sup_{\theta\in V\eta_{\theta_j}(\theta_j)} \frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(\theta) \leq \sup_{\theta\in V}\frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(\theta),$$

which implies $\hat{\theta} \in V$.

□

### 6.4 Proof of Theorem 3.2

We here only give an outline of the proof, a detailed proof can be found in the supplementary material.

*Proof.* By Taylor's expansion, for $j = 1, \ldots, 6$, there exists some $\theta_{(j)}$ between $\theta_0$ and $\hat{\theta}$ such that

$$0 = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \tilde{\ell}_t(\hat{\theta})}{\partial \theta_j} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \tilde{\ell}_t(\theta_0)}{\partial \theta_j} + \left( \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \tilde{\ell}_t(\theta_{(j)})}{\partial \theta_j \partial \theta^\intercal} \right) \sqrt{n}(\hat{\theta} - \theta_0).$$

The theorem follows if it can be proved that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \tilde{\ell}_t(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, G),$$

and

$$\frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \tilde{\ell}_t(\theta^*)}{\partial \theta \partial \theta^\intercal} \xrightarrow{p} -G,$$

for all $\theta^*$ between $\theta_0$ and $\hat{\theta}$.

To show these, we prove the following statements,

(S1). $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \ell_t(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, G)$.

(S2). $\|\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\frac{\partial \ell_t(\theta_0)}{\partial \theta} - \frac{\partial \tilde{\ell}_t(\theta_0)}{\partial \theta})\| \xrightarrow{p} 0$.

(S3). There exists a neighbourhood of $\theta_0$, $V(\theta_0)$, such that for all $i, j, k \in \{1, \ldots, 6\}$,

$$\mathrm{E}\left( \sup_{\theta \in V(\theta_0)} |\frac{\partial^3 \ell_t(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}| \right) < \infty.$$

(S4). For the neighbourhood $V(\theta_0)$ specified above,

$$\sup_{\theta \in V(\theta_0)} \| \frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial^2 \ell_t(\theta)}{\partial \theta \partial \theta^\intercal} - \frac{\partial^2 \tilde{\ell}_t(\theta)}{\partial \theta \partial \theta^\intercal} \right) \| \xrightarrow{p} 0.$$

(S5). $\frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \ell_t(\theta^*)}{\partial \theta \partial \theta^\intercal} \xrightarrow{a.s.} -G$, uniformly for all $\theta^*$ between $\theta_0$ and $\hat{\theta}$.

□

# References

Billingsley, P. (1999) *Convergence of Probability Measures.* Wiley-Interscience publication.

Blasques, F., Koopman, S. and Lucas, A. (2012) Stationarity and ergodicty of univariate generalized autoregressive score processes. *Tinbergen Institute discussion paper* .

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Cheng, X., Li, W. K., Yu, P. L. H., Zhou, X., Wang, C. and Lo, P. H. (2011) Modeling threshold conditional heteroscedasticity with regime-dependent skewness and kurtosis. *Computional Statistics and Data Analysis* **55**(9), 2590–2604.

Cox, D. R. (1981) Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics* **8**(2), 93–115.

Davis, R. and Liu, H. (2012) Theory and inference for a class of nonlinear models with application to time series of counts. *arXiv:1204.3915v1* .

Davis, R. A., Dunsmuir, W. T. M. and Streett, S. B. (2003) Observation-driven models for Poisson counts. *Biometrika* **90**(4), 777–790.

Diaconis, P. and Freedman, D. (1999) Iterated random functions. *SIAM Review* **41**(1), 45–76.

Douc, R., Doukhan, P. and Moulines, E. (2013) Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications* **123**, 2620–2647.

Doukhan, P., Fokianos, K. and Tjøstheim, D. (2012) On weak dependence conditions for poisson autoregressions. *Statistics & Probability Letters* **82**(5), 942–948.

Duflo, M. (1997) *Random Iterative Models.* Springer.

Ferland, R., Latour, A. and Oraichi, D. (2006) Integer-valued GARCH processes. *Journal of Time Series Analysis* **27**, 923–942.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009) Poisson autoregression. *Journal of the American Statistical Association* **104**(488), 1430–1439.

Fokianos, K. and Tjøstheim, D. (2011) Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102**, 563–578.

Fokianos, K. and Tjøstheim, D. (2012) Nonlinear poisson autoregression. *Annals of the Institute of Statistical Mathematics* **64**, 1205–1225.

Francq, C. and Zakoïan, J.-M. (2004) Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* **10**(4), 605–637.

Meyn, S. P. and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability.* Springer-Verlag, New York.

Neumann, M. (2011) Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17**(4), 1268–1284.

Tong, H. (1990) *Non-Linear Time Series: A Dynamical System Approach.* Oxford University Press.

Woodard, D. B., Matteson, D. S. and Henderson, S. G. (2011) Stationarity of generalized autoregressive moving average models. *Electronic Journal of Statistics* **5**, 800–828.

Wu, W. and Shao, X. (2004) Limit theorems for iterated random functions. *Journal of Applied Probability* **41**(2), 425–436.

Zucchini, W. and MacDonald, I. L. (2009) *Hidden Markov models for Time Series: an Introduction Using R.* Monographs on statistics and applied probability;110. CRC Press.

# Supplementary material

## Complementary for establishing the statements in the proof of Theorem 3.2

We write $\lambda_t$ as in Eq (5), then

$$\frac{\partial \ell_t}{\partial \theta} = (\frac{Y_t}{\lambda_t} - 1)\frac{\partial \lambda_t}{\partial \theta},$$

and

$$\frac{\partial \lambda_t}{\partial \theta} = \begin{pmatrix} \frac{\partial \lambda_t}{\partial \theta^{(1)}} \\ \\ \frac{\partial \lambda_t}{\partial \theta^{(2)}} \end{pmatrix}, \tag{15}$$

with

$$\frac{\partial \lambda_t}{\partial \theta^{(i)}} = \begin{pmatrix} 1 \\ \\ \lambda_{t-1} \\ \\ Y_{t-1} \end{pmatrix} 1\left\{Y_{t-1} \in R_i\right\} + a_{t-1}\frac{\partial \lambda_{t-1}}{\partial \theta^{(i)}} \quad (i = 1, 2).$$

The derivative in Eq (15) can be written in a compact form as

$$\frac{\partial \lambda_t}{\partial \theta} := \nu_{t-1} + a_{t-1}\frac{\partial \lambda_{t-1}}{\partial \theta} = \sum_{k\geq 1}(\prod_{j=1}^{k-1} a_{t-j})\nu_{t-k}.$$

By assumption $a_t \leq \max\left\{a_1, a_2\right\} = a_M < 1$, then

$$\frac{\partial \lambda_t}{\partial \theta} \leq \sum_k a_M^{k-1}\nu_{t-k}.$$

In particular, we have

$$\frac{\partial \lambda_t}{\partial d_i} = \sum_{k\geq 1}(\prod_{j=1}^{k-1} a_{t-j})1\left\{Y_{t-1} \in R_i\right\} \leq \sum_{k\geq 1} a_M^{k-1} \leq \frac{1}{1 - a_M}. \tag{16}$$

Writing $\lambda_t = \sum_{k\geq 1}(\prod_{j=1}^{k-1} a_{t-j})c_{t-k}$ with $c_t = d_t + b_t Y_t$, we have

$$\frac{\partial \lambda_t}{\partial b_i} = \sum_{k\geq 1}(\prod_{j=1}^{k-1} a_{t-j})\frac{\partial b_{t-k}}{\partial b_i} Y_{t-k} = \sum_{k\geq 1}(\prod_{j=1}^{k-1} a_{t-j}) 1\left\{Y_{t-k} \in R_i\right\} Y_{t-k},$$

which implies

$$\|\frac{\partial \lambda_t}{\partial b_i}\|_2 \leq \|Y_t\|_2 \sum_{k\geq 1} a_M^k. \tag{17}$$

Also,

$$\frac{\partial \lambda_t}{\partial a_i} = \sum_{k\geq 1} \frac{\partial(\prod_{j=1}^{k-1} a_{t-j})}{\partial a_i} c_{t-k} \leq \sum_{k\geq 1} \frac{k-1}{a_i}(\prod_{j=1}^{k-1} a_{t-j})c_{t-k},$$

implies

$$\mathrm{E}\left(\frac{\partial \lambda_t}{\partial a_i}\right) \leq \sum_{k\geq 1} \frac{k-1}{a_i} a_M^{k-1}(d_M + b_M \mathrm{E}(Y_t)) < \infty, \tag{18}$$

where $d_M = \max\left\{d_1, d_2\right\}, b_M = \max\left\{b_1, b_2\right\}$, and

$$\|\frac{\partial \lambda_t}{\partial a_i}\|_2 \leq \sum_{k\geq 1} \frac{k-1}{a_i} a_M^{k-1}(d_M + b_M\|Y_t\|_2) < \infty. \tag{19}$$

Note that

$$\mathrm{E}\left[\frac{\partial \ell_t(\theta_0)}{\partial \theta}\right] = \mathrm{E}\left[\left(\frac{Y_t}{\lambda_t} - 1\right)\frac{\partial \lambda_t}{\partial \theta}\right] = \mathrm{E}\left[\mathrm{E}\left(\frac{Y_t}{\lambda_t} - 1\right)\frac{\partial \lambda_t}{\partial \theta}|\mathcal{F}_{t-1}\right] = 0.$$

Since $\lambda_t$ is bounded from zero, $\lambda_t \geq d_0 = \min\left\{d_1, d_2\right\}$, with the results in

Eq (16), Eq (17), Eq (18), and Eq (19) we have

$$\begin{aligned}
\operatorname{var}\left[\frac{\partial \ell_t(\theta_0)}{\partial \theta}\right] =& \mathrm{E}\left[\left(\frac{Y_t}{\lambda_t}-1\right)^2\left(\frac{\partial \lambda_t}{\partial \theta}\right)\left(\frac{\partial \lambda_t}{\partial \theta}\right)^{\intercal}\right] \\
=& \mathrm{E}\left[\mathrm{E}\left\{\left(\frac{Y_t}{\lambda_t}-1\right)^2\left(\frac{\partial \lambda_t}{\partial \theta}\right)\left(\frac{\partial \lambda_t}{\partial \theta}\right)^{\intercal} \mid \mathcal{F}_{t-1}\right\}\right] \\
=& \mathrm{E}\left[\frac{1}{\lambda_t}\left(\frac{\partial \lambda_t}{\partial \theta}\right)\left(\frac{\partial \lambda_t}{\partial \theta}\right)^{\intercal}\right] \\
=& G < \infty.
\end{aligned}$$

It can be seen that $G$ is non-degenerate (cf. Francq and Zakoïan (2004)). Since $\{\partial \ell_t(\theta_0)/\partial\theta\}$ is a $L_4$ martingale difference, by the Cramér-Wold device and the central limit theorem in Theorem 18.1 of Billingsley (1999) we have the weak convergence,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial \ell_t(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, G).$$

Then we shall prove Statement (S2). To show this, note that for $i = 1, 2$,

$$\frac{\partial \tilde{\lambda}_t}{\partial d_i} = \sum_{k\geq 1}^{t-2}(\prod_{j=1}^{k-1} a_{t-j})1\{Y_{t-k} \in R_i\} + \prod_{j=1}^{k-1} a_{t-j}\frac{\partial \tilde{\lambda}_1}{\partial d_i}, \tag{20}$$

$$\frac{\partial \tilde{\lambda}_t}{\partial a_i} = \sum_{k=1}^{t-2}\frac{\partial(\prod_{j=1}^{k-1} a_{t-j})}{\partial a_i}c_{t-k} + \prod_{j=1}^{t-1} a_{t-j}\frac{\partial \tilde{\lambda}_1}{\partial a_i}, \tag{21}$$

$$\frac{\partial \tilde{\lambda}_t}{\partial b_i} = \sum_{k=1}^{t-2}(\prod_{j=1}^{k-1} a_{t-j})Y_{t-k}1\{Y_{t-k} \in R_i\} + \prod_{j=1}^{t-1} a_{t-j}\frac{\partial \tilde{\lambda}_1}{\partial b_i}. \tag{22}$$

Since $\partial\tilde{\lambda}_1/\partial\theta$ can be regarded as a fixed value, we have

$$\sup_{\theta\in\mathcal{D}}\|\frac{\partial \tilde{\lambda}_t}{\partial \theta} - \frac{\partial \lambda_t}{\partial \theta}\| \leq C\rho^t, a.s.$$

Note that we also have $|\lambda_t - \tilde{\lambda}_t| \leq C\rho^t$, which implies $|\frac{1}{\lambda_t} - \frac{1}{\tilde{\lambda}_t}| \leq C\rho^t$, for $\lambda_t$ and $\tilde{\lambda}_t$ are bounded from 0. Note that

$$\begin{aligned}\frac{\partial \ell_t(\theta_0)}{\partial \theta} - \frac{\partial \tilde{\ell}_t(\theta_0)}{\partial \theta} &= \left(\frac{Y_t}{\lambda_t(\theta_0)} - 1\right)\frac{\partial \lambda_t(\theta_0)}{\partial \theta} - \left(\frac{Y_t}{\tilde{\lambda}_t(\theta_0)} - 1\right)\frac{\partial \tilde{\lambda}_t(\theta_0)}{\partial \theta} \\ &= Y_t\left[\left(\frac{1}{\lambda_t} - \frac{1}{\tilde{\lambda}_t}\right)\frac{\partial \lambda_t}{\partial \theta} + \frac{1}{\tilde{\lambda}_t}\left(\frac{\partial \lambda_t}{\partial \theta} - \frac{\partial \tilde{\lambda}_t}{\partial \theta}\right)\right] - \left(\frac{\partial \lambda_t}{\partial \theta} - \frac{\partial \tilde{\lambda}_t}{\partial \theta}\right).\end{aligned}$$

Then it is readily seen that

$$\|\frac{\partial \ell_t(\theta_0)}{\partial \theta} - \frac{\partial \tilde{\ell}_t(\theta_0)}{\partial \theta}\| \leq C\rho^t\left[1 + Y_t\left(1 + \|\frac{\partial \lambda_t}{\partial \theta}\|\right)\right].$$

Note that $\mathrm{E}(Y_t\|\partial\lambda_t(\theta_0)/\partial\theta\|) < \infty$, then for any $\varepsilon > 0$,

$$\begin{aligned}\mathrm{pr}\left(\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(\frac{\partial \ell_t(\theta_0)}{\partial \theta} - \frac{\partial \tilde{\ell}_t(\theta_0)}{\partial \theta}\right)\| > \varepsilon\right) &\leq \frac{1}{\sqrt{n}\varepsilon}\sum_{t=1}^{n} C\rho^t\left[1 + \mathrm{E}(Y_t) + \mathrm{E}\left(\|Y_t\frac{\partial \lambda_t}{\partial \theta}\|\right)\right] \\ &\to 0, \text{ as } n \to \infty.\end{aligned}$$

Next we will prove Statement (S3). Through direct calculation, we obtain

$$\begin{aligned}\frac{\partial^3 \ell_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k} = &\left(-\frac{Y_t}{\lambda_t^2}\right)\left(\frac{\partial^2\lambda_t}{\partial\theta_i\partial\theta_j}\frac{\partial\lambda_t}{\partial\theta_k} + \frac{\partial^2\lambda_t}{\partial\theta_i\partial\theta_k}\frac{\partial\lambda_t}{\partial\theta_j} + \frac{\partial^2\lambda_t}{\partial\theta_j\partial\theta_k}\frac{\partial\lambda_t}{\partial\theta_i}\right) \\ &+ 2\frac{Y_t}{\lambda_t^3}\frac{\partial\lambda_t}{\partial\theta_i}\frac{\partial\lambda_t}{\partial\theta_j}\frac{\partial\lambda_t}{\partial\theta_k} + \left(\frac{Y_t}{\lambda_t} - 1\right)\frac{\partial^3\lambda_t}{\partial\theta_i\partial\theta_j\partial\theta_k}. \end{aligned} \tag{23}$$

Consider, for example, $\partial^3\ell_t(\theta)/\partial a_1^3$. Write $\lambda_t = \sum_k \prod_{j=1}^{k-1} a_{t-j}c_{t-k}$, then for $i = 1, 2, 3$,

$$\frac{\partial^i \lambda_t(\theta)}{\partial a_1^i} = \sum_{k\geq 1}\frac{\partial^i(\prod_{j=1}^{k-1} a_{t-j})}{\partial a_1^i}c_{t-k} \leq \sum_{k\geq 1}\frac{(k-1)\cdots(k-i)}{a_1^i}(\prod_{j=1}^{k-1} a_{t-j})c_{t-k}.$$

We may select $V(\theta_0)$ small enough such that $a_M = \sup_{\theta \in V(\theta_0)} \max\{a_1, a_2\} < 1$, and $a_m = \inf_{\theta \in V(\theta_0)} \min\{a_1, a_2\} > 0$, then

$$\frac{\partial^i \lambda_t(\theta)}{\partial a_1^i} \leq \sum_{k \geq 1} \frac{(k-1)\cdots(k-i)}{a_m^i} a_M^{k-1} c_{t-k} \quad (i = 1, 2, 3).$$

Recall that $c_t = d_t + a_t Y_t$, then it is easily seen that there exist constants $\zeta_{t,i} > 0$, such that $\sum_t \zeta_{t,i} < \infty$, and

$$\sup_{\theta \in V(\theta_0)} \frac{\partial^i \lambda_t(\theta)}{\partial a_1^i} \leq \zeta_{0,i} + \sum_{k \geq 1} \zeta_{k,i} Y_{t-k} := \mu_{t,i}.$$

From Eq (23), we have

$$\mathrm{E}\left( \sup_{\theta \in V(\theta_0)} |\frac{\partial^3 \ell_t(\tilde{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k}| \right) \leq \mathrm{E}\left[ 3\frac{Y_t}{d_m^2} \mu_{t,2}\mu_{t,1} + 2\frac{Y_t}{d_m^3} \mu_{t,1}^3 + \left( \frac{Y_t}{d_m} + 1 \right) \mu_{t,3} \right].$$

The expression on the right-hand-side of the inequality can be proved to be finite, if $\mu_{t,3} \in L_2$, $\mu_{t,1} \in L_6$, $\mu_{t,2} \in L_4$, which can be verified by Minkowski inequality and the fact that $Y_t \in L_p$, for all $p > 0$.

As for the second order derivative in Statement (S4), note that similar to the case for the first order derivative, we can prove

$$\sup_{\theta \in \boldsymbol{\Theta}} \| \frac{\partial^2 \lambda_t}{\partial \theta \partial \theta^\intercal} - \frac{\partial^2 \tilde{\lambda}_t}{\partial \theta \partial \theta^\intercal} \| \leq C\rho^t. \tag{24}$$

It is easily seen that

$$\frac{\partial^2 \ell_t}{\partial \theta \partial \theta^\intercal} = \left( \frac{Y_t}{\lambda_t} - 1 \right) \frac{\partial^2 \lambda_t}{\partial \theta \partial \theta^\intercal} - \frac{Y_t}{\lambda_t^2} \left( \frac{\partial \lambda_t}{\partial \theta} \right) \left( \frac{\partial \lambda_t}{\partial \theta} \right)^\intercal,$$

and

$$\mathrm{E}\left( \frac{\partial^2 \ell_t(\theta_0)}{\partial \theta \partial \theta^\intercal} \right) = -G.$$

Then

$$\begin{aligned}
&\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_k} - \frac{\partial^2 \tilde{\ell}_t}{\partial\theta_i\partial\theta_k} \\
=&Y_t\left[\left(\frac{1}{\lambda_t}-\frac{1}{\tilde{\lambda}_t}\right)\frac{\partial^2\lambda_t}{\partial\theta_i\partial\theta_k}+\frac{1}{\tilde{\lambda}_t}\left(\frac{\partial^2\lambda_t}{\partial\theta_i\partial\theta_k}-\frac{\partial^2\tilde{\lambda}_t}{\partial\theta_i\partial\theta_k}\right)+\left(\frac{1}{\lambda_t^2}-\frac{1}{\tilde{\lambda}_t^2}\right)\frac{\partial\lambda_t}{\partial\theta_i}\frac{\partial\lambda_t}{\partial\theta_k}\right. \\
&\left.+\frac{1}{\tilde{\lambda}_t^2}\left\{\frac{\partial\lambda_t}{\partial\theta_i}\left(\frac{\partial\lambda_t}{\partial\theta_j}-\frac{\partial\tilde{\lambda}_t}{\partial\theta_j}\right)+\frac{\partial\tilde{\lambda}_t}{\partial\theta_j}\left(\frac{\partial\lambda_t}{\partial\theta_i}-\frac{\partial\tilde{\lambda}_t}{\partial\theta_i}\right)\right\}\right]+\left(\frac{\partial^2\lambda_t}{\partial\theta_i\partial\theta_k}-\frac{\partial^2\tilde{\lambda}_t}{\partial\theta_i\partial\theta_k}\right).
\end{aligned}$$

Thus, we have

$$|\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_k} - \frac{\partial^2 \tilde{\ell}_t}{\partial\theta_i\partial\theta_k}| \leq C\left[1+Y_t\left(\frac{\partial^2\lambda_t}{\partial\theta_i\partial\theta_k}+\frac{\partial\lambda_t}{\partial\theta_i}\frac{\partial\lambda_t}{\partial\theta_k}+\frac{\partial\lambda_t}{\partial\theta_i}+\frac{\partial\lambda_t}{\partial\theta_k}\right)\right]\rho^t.$$

Let

$$\Gamma_t = \frac{\partial^2\lambda_t}{\partial\theta_i\partial\theta_k}+\frac{\partial\lambda_t}{\partial\theta_i}\frac{\partial\lambda_t}{\partial\theta_k}+\frac{\partial\lambda_t}{\partial\theta_i}+\frac{\partial\lambda_t}{\partial\theta_k},$$

then it can be seen that around a neighbourhood of $\theta_0$, without loss of generality, assuming the same $V(\theta_0)$, we have $\sup_{\theta\in V(\theta_0)} \mathrm{E}\left(\Gamma_t Y_t\right) < \infty$.

Similar as in the argument for Statement (S3), we can obtain the following by Markov inequality,

$$\sup_{\theta\in\mathbf{\Theta}} |\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_j} - \frac{\partial^2 \tilde{\ell}_t}{\partial\theta_i\partial\theta_j}\right)| \xrightarrow{p} 0. \tag{25}$$

Lastly, we prove Statement (S5). Recall that $\theta^*$ lies between $\theta_0$ and $\hat{\theta}$. Consider the Taylor expansion of the second-order derivatives of $\ell_t$ at $\theta_0$, we have

$$\frac{1}{n}\sum_t \frac{\partial^2\ell_t(\theta^*)}{\partial\theta_i\partial\theta_j} = \frac{1}{n}\sum_t \frac{\partial^2\ell_t(\theta_0)}{\partial\theta_i\partial\theta_j} + \frac{1}{n}\sum_t \frac{\partial^3\ell_t(\tilde{\theta})}{\partial\theta_i\partial\theta_j\partial\theta}(\theta^*-\theta_0),$$

for some $\tilde{\theta}$ between $\theta_0$ and $\theta^*$. Then the almost sure convergence of $\tilde{\theta}$ to $\theta_0$, the ergodic theorem in Corollary 2.4, and Statement (S3) imply that

$$\overline{\lim} \sup_{\theta \in V(\theta_0)} \| \frac{1}{n} \sum_t \frac{1}{n} \frac{\partial^3 \ell_t(\theta)}{\partial \theta_i \partial \theta_j \partial \theta} \| < \infty, a.s.$$

Then we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_t \frac{\partial^2 \ell_t(\theta^*)}{\partial \theta_i \partial \theta_j} = \lim_{n \to \infty} \frac{1}{n} \sum_t \frac{\partial^2 \ell_t(\theta_0)}{\partial \theta_i \partial \theta_j} = -G(i, j) \; a.s.$$

The proof is complete.