Ann Inst Stat Math (2012) 64:1205–1225
DOI 10.1007/s10463-012-0351-3

Nonlinear Poisson autoregression

Konstantinos Fokianos · Dag Tjøstheim

Received: 28 June 2010 / Revised: 18 October 2011 / Published online: 13 March 2012

Abstract We study statistical properties of a class of non-linear models for regression analysis of count time series. Under mild conditions, it is shown that a perturbed version of the model is geometrically ergodic and possesses moments of any order. This result turns out to be instrumental on deriving large sample properties of the maximum likelihood estimators of the regression parameters. The theory is illustrated with examples.

Keywords Geometric ergodicity · Link function · Maximum likelihood estimation · Perturbation · Smooth transition models

1 Introduction

Suppose that $\{Y_t\}$ is a time series of counts and let $\mathcal{F}_t^{Y,\lambda}$ be the σ-field generated by $\{Y_0, \ldots, Y_t, \lambda_0\}$, where λ_0 is the initial value of a Poisson intensity process $\{\lambda_t\}$. In Fokianos et al. (2009a,b) we studied the linear Poisson autoregressive model

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

for $t \geq 1$, and where the parameters d, a, b are assumed to be positive and λ_0, Y_0 are fixed.

K. Fokianos (✉)
Department of Mathematics, University of Cyprus, PO BOX 20537, 1678 Nicosia, Cyprus
e-mail: fokianos@ucy.ac.cy

D. Tjøstheim
Department of Mathematics, University of Bergen, Johannes Bruns gate 12, 5008 Bergen, Norway
e-mail: Dag.Tjostheim@math.uib.no

To study the probabilistic properties of (1), it is advantageous (cf. Fokianos et al. 2009a) to reformulate the model such that the sequence of independent Poisson drawings is expressed explicitly in terms of random variables. This strategy bears some analogy with the observational equation in a state space model, or like the defining equation of a GARCH model giving the relationship between the observations and the conditional variance. Towards this goal, for each time point t, a Poisson process $N_t(\cdot)$ of unit intensity is introduced. Then, the first part of (1) can be restated in terms of these Poisson processes by assuming that Y_t is equal to the number of events of $N_t(\lambda_t)$ of $N_t(\cdot)$ in the time interval $[0, \lambda_t]$. Let therefore $\{N_t(\cdot), t = 1, 2 \ldots\}$ be a sequence of independent Poisson processes of unit intensity and formulate model (1) as

$$Y_t = N_t(\lambda_t), \quad \lambda_t = d + a\lambda_{t-1} + bY_{t-1}, \tag{2}$$

for $t \geq 1$ and with Y_0, λ_0 fixed.

In this paper we will study the following nonlinear generalization of (2) (or equivalently of (1)):

$$Y_t = N_t(\lambda_t), \quad \lambda_t = f(\lambda_{t-1}) + b(Y_{t-1}), \tag{3}$$

for $t \geq 1$. In the above, $f(\cdot)$ and $b(\cdot)$ are known functions up to an unknown finite dimensional parameter vector. Moreover, both functions take values on the positive real line, that is, $f, b : R^+ \to R^+$ and, the initial values Y_0 and λ_0 are assumed to be fixed. It can be shown that the process $\{\lambda_t\}$ in (3) can be expressed as a function of past Y_t's, and λ_0 after repeated substitution. In other words, the hidden process $\{\lambda_t\}$ is determined by past functions of lagged responses, or equivalently, (3) belongs to the class of observation driven models in the sense of Cox (1981).

Model (2) is a special case of (3) upon defining $f(x) = d + ax$ and with a slight abuse of notation, $b(x) = bx$, with $d, a, b > 0$, and $x > 0$. The nonlinear autoregressive Poisson process (3) constitutes an analogy of ordinary nonlinear autoregressive models as treated, for example, in Tong (1990) and Fan and Yao (2003). With the exception of Fokianos et al. (2009a), Fokianos et al. (2009b), Neumann (2011), Franke (2010) and Doukhan et al. (2012) these types of models have not been considered before in the literature. Hence, our study enriches the class of models for count time series and provides simultaneously tools for inference and testing.

The recent contribution by Neumann (2011) proves, under a uniform contractivity condition, that the process $\{Y_t, \lambda_t\}$ is ergodic. However, the author recognizes the fact that $\{Y_t, \lambda_t\}$ and even the intensity process $\{\lambda_t\}$ alone are not strongly mixing in general. Moreover, only the second moment of $\{\lambda_t\}$ is shown to exist. These findings make it difficult to prove asymptotic normality of the maximum likelihood estimators of any finite dimensional parameter vector contained in $f(\cdot)$ and $b(\cdot)$ of (3), as it will be explained in Sect. 4. Possibly, Neumann's results combined with weak dependence results derived in Franke (2010) and Doukhan et al. (2012) could be used for this purpose, but it is far from straightforward, and we have chosen the alternative way based on Fokianos et al. (2009a). In that approach, asymptotic normality of the parameter estimates of a perturbed process is proved. Then asymptotic normality for the original parameter estimates is obtained by a limiting argument connecting the

two representations. This method of proof requires geometric ergodicity of the perturbed version of $\{\lambda_t\}$ (where the results by Neumann do not apply). The perturbation means that the traditional drift criterion for Markov chains can be used for proving ϕ-irreducibility. However, if the drift criterion is applied to the non perturbed model, then ϕ-irreducibility cannot be established. In fact, we are only able to show open set irreducibility and this is not sufficient to prove geometric ergodicity (cf. Fokianos et al. 2009b, Lemma A1).

Therefore, we study the ergodic properties of a perturbed version of (3), namely

$$\begin{aligned} Y_t^m &= N_t(\lambda_t^m), \quad \lambda_t^m = f(\lambda_{t-1}^m) + b(Y_{t-1}^m) + \varepsilon_{t,m} \\ \varepsilon_{t,m} &= c_m 1(Y_t^m = 1)U_t, \quad c_m > 0, \quad U_t \text{ iid } U[0,1], \end{aligned} \tag{4}$$

and where $\{Y_t^m\}$ can be identified with $\{Y_t\}$ as observations but not as random variables. The perturbation can be included in many other ways, and the results of this paper will still remain valid. For example, the indicator function $1(Y_t^m = 1)$ can be removed. For further motivation and details for the perturbation methodology we refer to Fokianos et al. (2009a,b).

In this contribution, we will outline a theory of inference for model (3) with asymptotic distributional results for the conditional maximum likelihood estimates of the unknown parameters contained in both functions $f(\cdot)$ and $b(\cdot)$. The perturbed version (4) will be instrumental in this analysis because geometric ergodicity is being proved for this version, and then asymptotics for the conditional maximum likelihood estimates in (3) are obtained by a limiting argument where we allow $c_m \downarrow 0$ in (4). Related literature regarding log-linear models for time series of counts include the works by Zeger and Qaqish (1988), Li (1994), MacDonald and Zucchini (1997). Brumback et al. (2000), Fahrmeir and Tutz (2001), Kedem and Fokianos (2002), Davis et al. (2003), Fokianos and Kedem (2004), Jung et al. (2006) and more recently Fokianos and Tjøstheim (2011).

An outline of the paper is as follows: In Sect. 2 we will briefly review the results on geometric ergodicity for the perturbed model (4) obtained in Fokianos et al. (2009a,b). In Sect. 3 we provide a link between the nonlinear model (3) that we are primarily interested in and its perturbed version (4). The main section of the paper is Sect. 4, where we derive the asymptotic theory of the conditional maximum likelihood estimates. Finally, in Sects. 5 and 6 we report some simulations and a real data example, respectively.

2 Geometric ergodicity

Geometric ergodicity of model (4) is important for proving asymptotic normality of the conditional maximum likelihood estimates of (4), which in turn is used to prove asymptotic normality of the estimates in (3) via a limiting process with $c_m \downarrow 0$ as $m \to \infty$. The ergodic problem was considered in Fokianos et al. (2009b), and we now briefly review and complement the results obtained there. Throughout the paper the following assumption—which in essence is identical to the assumption **NL** in Fokianos et al. (2009b)—on functions $f(\cdot)$ and $b(\cdot)$ will be used.

Assumption A1 (i) There exists a unique solution, denoted λ^*, of the equation $\lambda = f(\lambda)$.

(ii) With λ positive real and y a positive integer, $f(\lambda)$ is increasing in λ for $\lambda > \lambda^*$ and $b(y)$ is increasing in y such that $b(y) \geq \beta^* y$, $\beta^* > 0$ and with $b(0) = 0$.

(iii) For some $\alpha_2 > 0$, $|f(\lambda_2) - f(\lambda_1)| \leq \alpha_2 |\lambda_2 - \lambda_1|$ for all $\lambda_1, \lambda_2 \geq 0$.

(iv) For some $\beta_2 > 0$ such that $\alpha_2 + \beta_2 < 1$, $b(y_2) - b(y_1) \leq \beta_2(y_2 - y_1)$, $y_2 \geq y_1$.

The conditions (i) and (ii) are used in proving that $\{\lambda_t\}$ is open set irreducible (cf. Fokianos et al. 2009a). When this is combined with the perturbation in (4), ϕ-irreducibility is obtained. The added conditions (iii) and (iv) imply geometric ergodicity and are used in establishing the asymptotic proximity of λ_t, λ_t^m and Y_t, Y_t^m in Proposition 3. Note that geometric ergodicity of the perturbed process can be proved under weaker conditions. The uniformity in (iii) and (iv) is needed in the proof of Proposition 3.

An example of a model fulfilling assumption A1 is the following model, see also Gao et al. (2009):

$$f(\lambda) = d\frac{1}{(1+\lambda)^{\gamma}} + a\lambda \quad \text{and} \quad b(y) = by, \tag{5}$$

provided that all the parameters d, a, b, γ are positive such that

$$\sup_{\lambda \geq 0} \left| a - d\frac{\gamma}{(1+\lambda)^{\gamma+1}} \right| + b < 1. \tag{6}$$

The inclusion of the parameter γ introduces a nonlinear deviation, in the sense that small values of the parameter γ cause (5) to approach model (1). Moderate values of γ introduce a stronger deviation.

Another interesting example of a non-linear regression model for count time series analysis is given by the following specification:

$$f(\lambda) = d + (a + c\exp(-\gamma\lambda^2))\lambda \quad \text{and} \quad b(y) = by, \tag{7}$$

where d, a, c, γ are positive parameters. It can be seen from (3) that $\lambda_t \geq d$. Moreover,

$$\frac{\partial f}{\partial \lambda} = a + ce^{-\gamma\lambda^2}(1 - 2\gamma\lambda^2)$$

and to satisfy A1 we must have

$$\sup_{\lambda \geq d} |a + ce^{-\gamma\lambda^2}(1 - 2\gamma\lambda^2)| + b < 1. \tag{8}$$

The above model parallels the structure of the traditional exponential autoregressive model, see Haggan and Ozaki (1981). In Fokianos et al. (2009a,b) model (7) was studied for the case $d = 0$. Several other examples are provided by the class of smooth transition autoregressive models of which the exponential autoregressive model is a special case (cf. Teräsvirta et al. 2010). We have the following result for the general perturbed nonlinear model (4), see Fokianos et al. (2009b, Prop. 2.3):

Proposition 1 *Consider the perturbed model* (4) *and suppose that assumption A1 holds. Then, the process* $\{\lambda_t^m, t \geq 0\}$ *is a geometrically ergodic Markov chain with finite moments of order k, for an arbitrary k.*

Proof The proof is essentially given in Fokianos et al. (2009b, Prop. 2.3), except that we did not provide the details of the aperiodicity part. Hence, referring to the proof of aperiodicity in the linear case of Fokianos et al. (2009b, Prop. 2.1), let λ^* be the fixed point of $f(\cdot)$ defined in condition A1 (i). Consider the small set $C = [\lambda^*, K]$. Note that for the ϕ-measure defined in the proof of Fokianos et al. (2009b, Prop. 2.1), we have $\phi(C) > 0$, and let $\lambda_{t-1}^m = \lambda \in C$. Then $\lambda_t^m = f(\lambda) + b(Y_{t-1}^m) + \varepsilon_{t,m}$. If $Y_{t-1}^m = 0$, then $\lambda_t^m = f(\lambda^*) + f(\lambda) - f(\lambda^*) = \lambda^* + f(\lambda) - f(\lambda^*) \geq \lambda^*$, since f is non-decreasing. On the other hand, $\lambda_t^m - \lambda = f(\lambda_{t-1}^m) - \lambda = f(\lambda^*) + f(\lambda) - f(\lambda^*) - \lambda \leq \lambda^* - \lambda + \alpha_2(\lambda - \lambda^*) = (1-\alpha_2)(\lambda^* - \lambda) \leq 0$, and the rest of the proof is as in Fokianos et al. (2009b). □

Consider again the defining equation (4) for the perturbed version of the nonlinear model (3). The following proposition shows that the joint trivariate process $(Y_t^m, U_t, \lambda_t^m)$ is $V_{(Y,U,\lambda}$-geometrically ergodic with $V_{Y,U,\lambda}(Y, U, \lambda) = 1 + Y^k + \lambda^k + U^k$, see Meyn and Tweedie (1993, p. 355) and Meitz and Saikonnen (2008).

Proposition 2 *Consider the perturbed model* (4) *and suppose that assumption A1 holds.Then the process* $\{(Y_t^m, U_t, \lambda_t^m), t \geq 0\}$ *is a* $V_{(Y,U,\lambda)}$*-geometrically ergodic Markov chain with* $V_{Y,U,\lambda}(Y, U, \lambda) = 1 + Y^k + \lambda^k + U^k$.

Proof This is proved as in Fokianos et al. (2009b, Prop. 2.2). The first part of the proof of that proposition can be adopted with virtually no changes to the present situation, since the conditioning arguments do not depend on differences in model structure between the perturbed version of model (2) and model (4). In the last part, where existence of a k-th order moment is shown, the same technique is used, just combined with the last part of Fokianos et al. (2009b, Prop. 2.3) for the evaluation of $E[\{b(Y_{t-1}^m)\}^k]$. □

We have the following results:

Corollary 1 *Consider the perturbed version of model* (5)

$$Y_t^m = N_t(\lambda_t^m), \quad \lambda_t^m = d\frac{1}{(1+\lambda_{t-1}^m)^\gamma} + a\lambda_{t-1}^m + bY_{t-1}^m + \varepsilon_{t,m},$$

with the same notation as in (4) *and suppose that* (6) *holds true. Then the process* $\{(Y_t^m, U_t, \lambda_t^m), t \geq 0\}$ *is* $V_{(Y,U,\lambda)}$*-geometrically ergodic Markov chain with* $V_{Y,U,\lambda}(Y, U, \lambda) = 1 + Y^k + \lambda^k + U^k$.

Corollary 2 *Consider the perturbed version of model* (7)

$$Y_t^m = N_t(\lambda_t^m), \quad \lambda_t^m = d + (a + c\exp(-\gamma(\lambda_{t-1}^m)^2))\lambda_{t-1}^m + bY_{t-1}^m + \varepsilon_{t,m},$$

with the same notation as in (4) *and suppose that* (8) *holds true. Then the process* $\{(Y_t^m, U_t, \lambda_t^m), t \geq 0\}$ *is* $V_{(Y,U,\lambda)}$*-geometrically ergodic Markov chain with* $V_{Y,U,\lambda}(Y, U, \lambda) = 1 + Y^k + \lambda^k + U^k$.

3 The limit of the perturbed version

To be able to carry through the likelihood arguments of the next section we need to establish that the models (3) and (4) are close as $c_m \downarrow 0$ in model (4). The following results establish this connection:

Proposition 3 *Suppose that* $\{Y_t, \lambda_t\}$ *and* $\{Y_t^m, \lambda_t^m\}$ *are defined by models* (3) *and* (4), *respectively. If condition A1 holds, then as* $c_m \downarrow 0$ *fast enough and* $\lambda_0^m = \lambda_0$, *the following results hold true:*

(i) $E|\lambda_t^m - \lambda_t| \le \delta_{1,m}$,
(ii) $E|(\lambda_t^m - \lambda_t)^2 \le \delta_{2,m}$,
(iii) $E|Y_t^m - Y_t| \le \delta_{3,m}$,
(iv) $E(Y_t^m - Y_t)^2 \le \delta_{4,m}$,
(v) *Almost surely* $|\lambda_t^m - \lambda_t| \to 0$ *and* $|Y_t^m - Y_t| \to 0$.

The sequences $\delta_{i,m}$ *can be chosen to be independent of* t *and* $\delta_{i,m} \to 0$ *as* $m \to \infty$, *for* $i = 1, \ldots, 4$.

Proof We first prove (i) and the first part of (v). By subtracting model (3) from model (4), we obtain that

$$\lambda_t^m - \lambda_t = f(\lambda_{t-1}^m) - f(\lambda_{t-1}) + b(Y_{t-1}^m) - b(Y_{t-1}) + \varepsilon_{t,m}.$$

Now, we use the fact that $\{Y_t^m\}$ and $\{Y_t\}$ are generated by the same sequence of independent Poisson processes $\{N_t(\cdot)\}$ of unit intensity. In other words, $Y_t^m = N_t(\lambda_t^m)$ is the number of events for $N_t(\cdot)$ in the stochastic time interval $[0, \lambda_t^m]$ and similarly for $Y_t = N_t(\lambda_t)$. Assume first that $\lambda_{t-1}^m \ge \lambda_{t-1}$. This implies that $Y_{t-1}^m \ge Y_{t-1}$ and $Y_{t-1}^m = \Delta Y_{t-1}^m + Y_{t-1}$, where ΔY_{t-1}^m is independent of Y_{t-1} given λ_{t-1} and λ_{t-1}^m. Then A1 (iii) yields

$$E|f(\lambda_{t-1}^m) - f(\lambda_{t-1}| \le \alpha_2 E|\lambda_{t-1}^m - \lambda_{t-1}|.$$

Similarly, A1 (iv) shows that

$$\begin{aligned} E|b(Y_{t-1}^m) - b(Y_{t-1})| &\le \beta_2 E|Y_{t-1}^m - Y_{t-1}| = \beta_2 E[E(|\Delta Y_{t-1}^m|\ |\lambda_{t-1}, \lambda_{t-1}^m)] \\ &= \beta_2 E|\lambda_{t-1}^m - \lambda_{t-1}|. \end{aligned}$$

Therefore, we have that

$$E|\lambda_t^m - \lambda_t| \le (\alpha_2 + \beta_2) E|\lambda_{t-1}^m - \lambda_{t-1}| + E|\varepsilon_{t,m}|. \tag{9}$$

By A1 (iv), $\alpha_2 + \beta_2 < 1$. Moreover, $E|\varepsilon_{t,m}| \le c_m \to 0$, and since $\lambda_0^m = \lambda_0$, it follows that $E|\lambda_t^m - \lambda_t| \to 0$ as $m \to \infty$. Because the rate at which $c_m \downarrow 0$ is at our disposal, we can prove that $\lambda_t^m - \lambda_t \to 0$ almost surely, and this completes the proof of (i) and the first part of (v).

Next, we note that $E(\lambda_t^k) < \infty$. This result can be proved along the lines of the proof of $E((\lambda_t^m)^k) < \infty$ (cf. Fokianos et al. 2009b, Prop. 2.3). We point out that the

property of ϕ-irreducibility is not required in that proof; the only conditions that are needed are $\alpha_2 + \beta_2 < 1$ and the Markov property. Then (ii) follows from the first part of (v) and Lebesgue dominated convergence because $E(\lambda_t^m - \lambda_t)^2$ is bounded independently of m.

Since $Y_t^m \geq Y_t$ if and only if $\lambda_t^m \geq \lambda_t$, item (iii) follows from $E(Y_t^m - Y_t) = E(E(Y_t^m - Y_t)|\lambda_t^m, \lambda_t)) = E(\lambda_t^m - \lambda_t)$ and (i). From this also follows the last part of (v), because $E(Y_t^m - Y_t)^2 = E(E(\Delta Y_t^m)^2|\lambda_t, \lambda_t^m)) = E(\lambda_t^m - \lambda_t)^2 + E^2(\lambda_t^m - \lambda_t)$, and (iv) follows from (i) and (ii). (Alternatively, one could use the last part of (v) and Lebesgue dominated convergence). By going through the proof step by step, it is seen that identical arguments can be used for the case $\lambda_{t-1} > \lambda_{t-1}^m$. This completes the proof. □

4 Likelihood inference

The principles of likelihood inference in the linear case have been studied extensively in Fokianos et al. (2009b). Because we retain the conditional Poisson assumption also in the nonlinear situation, there are considerable structural similarities. We assume that $f(\cdot)$ and $b(\cdot)$ in both (3) and (4) are known up to some unknown finite dimensional parameters, $\boldsymbol{\theta}_1 = (\theta_{11}, \ldots, \theta_{1p})'$ and $\boldsymbol{\theta}_2 = (\theta_{21}, \ldots, \theta_{2q})'$ say, which make up the unknown parameter vector $\boldsymbol{\theta}$.

The conditional likelihood function for $\boldsymbol{\theta}$ based on (3) and given the starting value λ_0 in terms of the observations $Y_1, \ldots, Y_n$ is given by

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

and the corresponding log-likelihood (omitting an unimportant constant) and score functions are given by

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} (Y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})), \tag{10}$$

and

$$\mathbf{S}_n(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta})}, \tag{11}$$

respectively. The solution of the equation $\mathbf{S}_n(\boldsymbol{\theta}) = 0$, provided that it exists, yields the maximum conditional likelihood estimator of $\boldsymbol{\theta}$, which is denoted by $\widehat{\boldsymbol{\theta}}$. The Hessian matrix for model (3) is obtained by differentiating the score function. That is

$$\begin{aligned} \mathbf{H}_n(\boldsymbol{\theta}) = &- \sum_{t=1}^{n} \frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{t=1}^{n} \frac{Y_t}{\lambda_t^2(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \\ &- \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial^2 \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}. \end{aligned} \tag{12}$$

Similarly for the perturbed system (4), it is assumed (artificially) that both $Y_1^m, \ldots Y_n^m$ and $U_1, \ldots, U_n$ are observed. In other words, we obtain

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

by the Poisson assumption and the asserted independence of U_t from $(Y_{t-1}^m, \lambda_{t-1}^m)$ with $f_u(\cdot)$ denoting the uniform density. Moreover,

$$l^m(\boldsymbol{\theta}) = \sum_{t=1}^{n} (Y_t^m \log \lambda_t^m(\boldsymbol{\theta}) - \lambda_t^m(\boldsymbol{\theta})) + \sum_{t=1}^{n} \log f_u(U_t),$$

$$\mathbf{S}_n^m(\boldsymbol{\theta}) = \frac{\partial l^m(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{Y_t^m}{\lambda_t^m(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t^m(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and

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

Recall the notation introduced in (4) and denote by $\mathcal{F}_t$ the σ-algebra generated by $\{U_{k+1}, N_k, k \leq t\}$. The proof of asymptotic normality of the conditional maximum likelihood estimates in Fokianos et al. (2009b) uses a martingale approach where the score has martingale difference terms, so that with $Z_t^m = (Y_t^m/\lambda_t^m - 1)$, we have $E(Z_t^m|\mathcal{F}_{t-1}) = 0$ and $E((Z_t^m)^2|\mathcal{F}_{t-1}) = 1/\lambda_t^m$, under the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This remains true in our more general case. But the evaluation of derivatives needed in the proof requires new regularity conditions. First note that the following is true:

$$\begin{aligned} \frac{\partial \lambda_t}{\partial \boldsymbol{\theta}_1} &= \frac{\partial f(\lambda_{t-1}, \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} + \frac{\partial f(\lambda_{t-1}, \boldsymbol{\theta}_1)}{\partial \lambda} \frac{\partial \lambda_{t-1}}{\partial \boldsymbol{\theta}_1}, \\ \frac{\partial \lambda_t}{\partial \boldsymbol{\theta}_2} &= \frac{\partial b(Y_{t-1}, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} + \frac{\partial f(\lambda_{t-1}, \boldsymbol{\theta}_1)}{\partial \lambda} \frac{\partial \lambda_{t-1}}{\partial \boldsymbol{\theta}_2}, \end{aligned} \tag{13}$$

and similarly for the perturbed system,

$$\begin{aligned} \frac{\partial \lambda_t^m}{\partial \boldsymbol{\theta}_1} &= \frac{\partial f(\lambda_{t-1}^m, \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} + \frac{\partial f(\lambda_{t-1}^m, \boldsymbol{\theta}_1)}{\partial \lambda} \frac{\partial \lambda_{t-1}^m}{\partial \boldsymbol{\theta}_1}, \\ \frac{\partial \lambda_t^m}{\partial \boldsymbol{\theta}_2} &= \frac{\partial b(Y_{t-1}^m, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} + \frac{\partial f(\lambda_{t-1}^m, \boldsymbol{\theta}_1)}{\partial \lambda} \frac{\partial \lambda_{t-1}^m}{\partial \boldsymbol{\theta}_2}. \end{aligned} \tag{14}$$

The above expressions show the importance of Proposition 2 for studying the large sample properties of the conditional maximum likelihood estimator. It is well known

that an asymptotic theory can be developed upon proving that the score evaluated at the true parameter is a martingale which converges to a normal random variable and the Hessian matrix has to converge in probability to a finite limit. In addition, the third derivatives need to be bounded by a sequence which converges in probability. Proposition 2 guarantees that all the above conditions are met for the perturbed model (4) when likelihood inference is based on the log-likelihood function $l^m(\boldsymbol{\theta})$. In particular, existence of moments guarantees the application of central limit theorem for martingales and ensures bounds for all necessary moments. In addition, geometric ergodicity assures that the Hessian matrix has a limit. Lemma 3 shows that the two models are close and this in turn implies that the large sample theory which is developed below is valid for model (3).

The following regularity conditions are assumed to hold: The parameter spaces for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ will be denoted by $\boldsymbol{\Theta_1}$ and $\boldsymbol{\Theta_2}$, respectively. We denote by $\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20}$, the true values of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. The notation $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}'_{10}, \boldsymbol{\theta}'_{20})'$ denotes the true value of $\boldsymbol{\theta}$.

A2 The parameter $\boldsymbol{\theta}$ belongs to an open set $\boldsymbol{\Theta}$ which is composed from $\boldsymbol{\Theta_1}$ and $\boldsymbol{\Theta_2}$. In addition, the function $f(\cdot)$ satisfies

$$f(\lambda, \boldsymbol{\theta}_1) \geq c_1 > 0,$$

for some positive constant c_1, $\boldsymbol{\theta_1} \in \boldsymbol{\Theta_1}$, $\lambda \geq 0$.

A3 The components of $\partial f/\partial \boldsymbol{\theta}_1$ and $\partial b/\partial \boldsymbol{\theta}_2$ are not linearly dependent; i.e., there do not exist a vectors $\mathbf{v}_i \neq \mathbf{0}, i = 1, 2$ so that $\mathbf{v_1}'\partial f/\partial \boldsymbol{\theta}_1 + \mathbf{v_2}'\partial b/\partial \boldsymbol{\theta}_2 = 0$.

A4 The functions $f(\cdot)$ and $b(\cdot)$ are four times continuously differentiable with respect to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, and all their derivatives are bounded. Moreover, $f(\cdot)$ is four times continuously differentiable with respect to $\boldsymbol{\theta}_1$ and λ and such that all of such derivatives are bounded by a linear function of λ, so that, for example, $||\partial^4 f/\partial \boldsymbol{\theta}_1 \partial \lambda^3|| \leq C_3\lambda$ for a non-negative C_3.

A5 The following identifiability condition holds:
- (a) $f(\lambda_{t-1}, \boldsymbol{\theta}_1) = f(\lambda_{t-1}, \boldsymbol{\theta}_{10})$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ and
- (b) $b(Y_{t-1}, \boldsymbol{\theta}_2) = b(Y_{t-1}, \boldsymbol{\theta}_{20})$ implies $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$.

Conditions A1–A5 are relatively mild. They are satisfied for the linear model (1), for models (5) and (7) under the corresponding stated condition for these examples in Sect. 2 and for similar smooth transition autoregressive models. In particular, condition A5 is equivalent to $\boldsymbol{\theta}_0$ being a locally unique asymptotic maximizer of the log-likelihood function (10); see Berkes et al. (2003, Theorem 2.3) and Francq and Zakoïan (2004, Assumption A4 and Remark 2.4). We are now ready to formulate our main theorem:

Theorem 1 *Consider model* (3) *and assume that the conditions* A1–A5 *are fulfilled. Then, there exists an open neighborhood* $O = O(\boldsymbol{\theta}_0)$ *of the true value* $\boldsymbol{\theta}_0$*, such that the probability that a locally unique maximum conditional likelihood estimator exists converge to one, as* $n \to \infty$*. Moreover, there exists a sequence of maximum conditional likelihood estimators* $\widehat{\boldsymbol{\theta}}$ *which is consistent and asymptotically normal,*

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

where the matrix $\mathbf{G}$ *is given by*

$$\mathbf{G}(\boldsymbol{\theta}) = E\left(\frac{1}{\lambda_t}\left(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}}\right)'\right). \tag{15}$$

A consistent estimator of $\mathbf{G}$ *is given by* $\mathbf{G}_n(\widehat{\boldsymbol{\theta}})/n$, *where*

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} Var\left[\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \mathcal{F}_{t-1}\right] = \sum_{t=1}^{n} \frac{1}{\lambda_t(\boldsymbol{\theta})}\left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)',$$

where $\mathcal{F}_t$ *is the* σ*-algebra generated by* $\{U_{k+1}, N_k, k \leq t\}$.

Remark 1 The expression in (15) requires the process $\{\lambda_t\}$ to be stationary. Strictly speaking, $\{\lambda_t\}$ will not be stationary if it is initiated with λ_0 fixed. However, the geometric ergodicity of the perturbed process $\{\lambda_t^m\}$ and the fact that $\lambda_t^m \to \lambda_t$ almost surely imply that the limit, as $t \to \infty$, in the right-hand side of (15) exists. It is this limit which by a slight abuse of notation is denoted by $\mathbf{G}(\theta)$. The same notation will be used in the sequel, and it was also used in Fokianos et al. (2009a) and Fokianos and Tjøstheim (2011).

The basic technique of proving the above theorem follows the arguments given in the proof of Fokianos et al. (2009b, Thm 3.1). It turns out that to study the asymptotic properties of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ for model (3), we need to derive and use the asymptotic properties of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}^m$ for the perturbed system (4). The main tool in connecting $\widehat{\boldsymbol{\theta}}$ to $\widehat{\boldsymbol{\theta}}^m$ is Brockwell and Davis (1991, Prop. 6.3.9). Accordingly, we first show that $\widehat{\boldsymbol{\theta}}^m$ is asymptotically normal, where the geometric ergodicity of the perturbed system is employed to derive the large sample results. Next, it is shown that the score function, the information matrix and the third derivative of the perturbed likelihood tend to the corresponding quantities for the unperturbed likelihood function, which in turn allows the application of Brockwell and Davis (1991, Prop. 6.3.9). This was proved in a series of lemmas for the case of the linear model (1), see Fokianos et al. (2009b, Lemmas 3.1–3.4). In the case of non-linear model (3) these results are restated below to complement our presentation. In addition, we give a proof for the existence and consistency of $\widehat{\boldsymbol{\theta}}$. All proofs are postponed to the Appendix.

Lemma 1 *Define the matrices*

$$\boldsymbol{G}^m(\boldsymbol{\theta}) = E\left(\frac{1}{\lambda_t^m}\left(\frac{\partial \lambda_t^m}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \lambda_t^m}{\partial \boldsymbol{\theta}}\right)'\right) \quad \textit{and} \quad \boldsymbol{G}(\boldsymbol{\theta}) = E\left(\frac{1}{\lambda_t}\left(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}}\right)'\right).$$

Under the assumptions of Theorem 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 2 *Under the assumptions of Theorem* 1, *the score functions defined by* (11) *and its perturbed counterpart 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. $\lim_{m\to\infty} \limsup_{n\to\infty} P(||\boldsymbol{S}_n^m - \boldsymbol{S}_n|| > \varepsilon\sqrt{n}) = 0$, *for every* $\varepsilon > 0$.

Lemma 3 *Under the assumptions of Theorem 1, the Hessian matrices defined by (12) and its perturbed counterpart 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. $\boldsymbol{G}^m \to \boldsymbol{G}$, *as* $m \to \infty$,
3. $\lim_{m\to\infty} \limsup_{n\to\infty} P(\|\boldsymbol{H}_n^m - \boldsymbol{H}_n\| > \varepsilon n) = 0$, *for every* $\varepsilon > 0$.

Lemma 4 *It holds under the assumptions of Theorem 1, with the neighborhood* $O(\boldsymbol{\theta}_0)$, *that*

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

Define correspondingly M_n^m *in terms of* Y_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. $\lim_{m\to\infty} \limsup_{n\to\infty} P(|M_n^m - M_n| > \varepsilon n) = 0$, *for every* $\varepsilon > 0$.

5 Examples

As an example of a non-linear model for count time series, we study (5) with γ known and fixed. The estimation results can be extended to the case of unknown γ along the lines of Fokianos et al. (2009b, Sec. 4.2). For real data, the parameter γ is unknown and this case is treated in Sect. 6. For model (5), $\boldsymbol{\theta}_1 = (d, a)$ and $\boldsymbol{\theta}_2 = b$. Recall that in Sect. 2, condition (6) was stated so that A1 is satisfied. The derivatives of $f(\cdot)$ and $b(\cdot)$ with respect to the parameters are given by

$$\frac{\partial f}{\partial d} = \frac{1}{(1+\lambda)^\gamma}, \quad \frac{\partial f}{\partial a} = \lambda, \frac{\partial b(y,b)}{\partial b} = y,$$

and it is seen that all of these derivatives satisfy assumption A4. Therefore, we should have a central limit theorem for the maximum likelihood estimates when (6) holds.

For the above model the score equations are given by (11) with recursions (13) satisfying the following:

$$\begin{aligned}
\frac{\partial\lambda_t}{\partial d} &= a\frac{\partial\lambda_{t-1}}{\partial d} + \frac{1}{(1+\lambda_{t-1})^\gamma} - \frac{d\gamma}{(1+\lambda_{t-1})^{\gamma+1}}\frac{\partial\lambda_{t-1}}{\partial d},\\
\frac{\partial\lambda_t}{\partial a} &= \lambda_{t-1} + a\frac{\partial\lambda_{t-1}}{\partial a} - \frac{d\gamma}{(1+\lambda_{t-1})^{\gamma+1}}\frac{\partial\lambda_{t-1}}{\partial a},\\
\frac{\partial\lambda_t}{\partial b} &= a\frac{\partial\lambda_{t-1}}{\partial b} - \frac{d\gamma}{(1+\lambda_{t-1})^{\gamma+1}}\frac{\partial\lambda_{t-1}}{\partial b} + Y_{t-1}.
\end{aligned}$$

Table 1 Maximum likelihood estimation results from 1000 simulations for model (5) with $\gamma = 1$ and for different parameter values and sample sizes

Parameter values			Estimators and SE			Sample size
a	d	b	$\hat{a}$	$\hat{d}$	$\hat{b}$	
0.30	1	0.40	0.2914 (0.0798)	1.0240 (0.2038)	0.3976 (0.0492)	500
			0.2974 (0.0537)	1.0107 (0.1444)	0.3994 (0.0326)	1000
0.30	0.25	0.40	0.2894 (0.0905)	0.2592 (0.0583)	0.3974 (0.0556)	500
			0.2996 (0.0632)	0.2528 (0.0423)	0.3956 (0.0393)	1000
0.50	2	0.40	0.4923 (0.0536)	2.1897 (0.5113)	0.3981 (0.0427)	500
			0.4947 (0.0390)	2.0983 (0.3781)	0.4012 (0.0304)	1000

Table 2 Maximum likelihood estimation results from 1000 simulations for model (5) with $\gamma = 2$ and for different parameter values and sample sizes

Parameter values			Estimators and SE			Sample size
a	d	b	$\hat{a}$	$\hat{d}$	$\hat{b}$	
0.20	1	0.50	0.2016 (0.0558)	1.0047 (0.1796)	0.4965 (0.0489)	500
			0.2001 (0.0411)	1.0008 (0.1263)	0.4972 (0.0340)	1000
0.20	0.25	0.50	0.1985 (0.0701)	0.2534 (0.0525)	0.4935 (0.0594)	500
			0.1985 (0.0506)	0.2531 (0.0361)	0.4978 (0.0438)	1000

We illustrate the asymptotic normality of the maximum likelihood estimators for model (5) by presenting some limited simulation results. In all cases considered, condition (6) is satisfied. Tables 1 and 2 report results from 1000 simulations for model (5) with $\gamma = 1$ and $\gamma = 2$, respectively. Note that for the data example to be discussed in Sect. 6, we have obtained that $\hat{\gamma} \approx 2$. To fit the model, we optimize the log-likelihood function (10) by a quasi-Newton method. Data from model (5) are generated for different parameter configurations and sample sizes and estimation is implemented by discarding the first 200 observations. Tables 1 and 2 report the estimates of the parameters by averaging out all the simulation output. In addition, the standard error of the estimators—in parentheses—is reported. In all cases, we note that the maximum likelihood estimators are consistent and their standard error decreases, as the sample size increases. Furthermore, the asymptotic normality is supported by Fig. 1 which shows qq-plots of the sampling distribution of the maximum likelihood estimators. The plots illustrate adequacy of the asymptotic normal distribution.

Turning now to model (7) with γ known and positive, we note that $\boldsymbol{\theta}_1 = (d, a, c)$ and $\boldsymbol{\theta}_2 = b$. Recall that conditions for A1 to be satisfied were given in Sect. 2. The derivatives of $f(\cdot)$ and $b(\cdot)$ with respect to the parameters are given by

$$\frac{\partial f}{\partial d} = 1, \quad \frac{\partial f}{\partial a} = \lambda, \quad \frac{\partial f}{\partial c} = e^{-\gamma\lambda^2}\lambda, \quad \text{and} \quad \frac{\partial b(y,b)}{\partial b} = y,$$

and it is seen that all of these derivatives satisfy assumption A4.

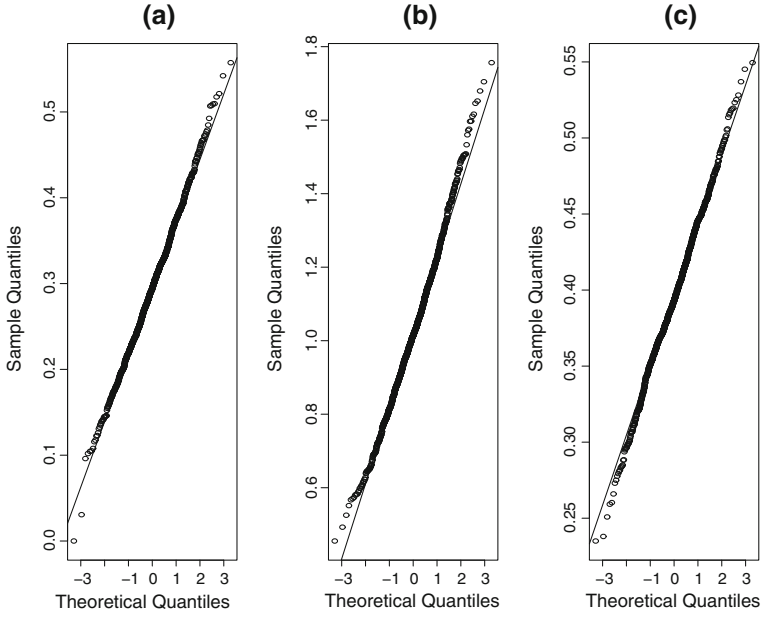

Fig. 1 QQplots of the sampling distribution of **a** $\hat{a}$, **b** $\hat{d}$ and **c** $\hat{b}$ for the non-linear model (5) when the true values are $a = 0.30$, $d = 1$ and $b = 0.40$. The results are based on 500 observations and 1000 simulation runs

We conclude that we should have a central limit theorem for the maximum likelihood estimates under the stated conditions. Even though we do not provide simulation results for model (7) we report that the introduction of the constant term d makes the estimation problem more challenging compared to the case $d = 0$, see Fokianos et al. (2009b). More specifically, estimation of the parameter c is more complicated. This can be explained by the fact that $\exp(-\gamma d^2)$ typically assumes small values even for small values of the parameter d. Therefore, the effect of parameter c cannot be identified since it is multiplied by a small number. When $d \sim 0$, then the parameter c is approximated more accurately but there appears to be a problem with estimation of d. In this case, constrained estimation shows that the asymptotic distribution of d has a positive mass at zero. These type of issues, even for the case of ordinary exponential autoregressive models, have been raised by several authors; see Chen et al. (2010) for a recent contribution.

6 Data analysis

We illustrate the application of the non-linear model (5) to real mortality data and we compare it to the linear count time series model (1). Figure 2 shows a time series plot of the daily number of deaths in Evora, Portugal, starting from 1st of January 1996 and ending at 31st of December 2007. The sample mean of the series is 6.119 and the variance is equal to 7.483, that is, the data are overdispersed. The plot illustrates that

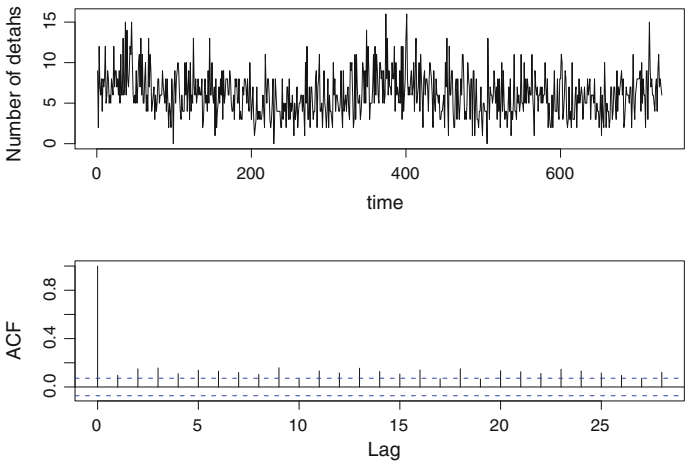

Fig. 2 Mortality data and their autocorrelation function

there is some type of periodicity in these data and the autocorrelation function is still significant for relatively large lag values.

First, we fit linear model (1) to these data. The results are given by $\hat{d} = 0.711$ (0.225), $\hat{a} = 0.762$ (0.051) and $\hat{b} = 0.122$ (0.024), where in parentheses are the standard errors of the estimators obtained by inversion of the information matrix, see Theorem 1. Note that the inclusion of the feedback mechanism is significant since the value of $\hat{a}$ is large when compared with its standard error. We also consider the Pearson residuals which are defined by $e_t = (Y_t - \lambda_t)/\sqrt{\lambda_t}$ and form a white noise sequence with constant variance, see Kedem and Fokianos (2002, Sec. 1.6.3). The sequence $\{e_t\}$ is calculated by substituting λ_t with $\lambda_t(\hat{\boldsymbol{\theta}})$. The mean square error of the Pearson residuals is defined by $\sum_{t=1}^{N} e_t^2/(N-p)$, where p denotes the number of estimated parameters.

For the mortality data and for the linear model application, the mean square error of Pearson residuals is equal to 1.150. The sequence of residuals and their cumulative periodogram plot are shown in Fig. 3. The bottom plot of this figure clearly indicates that the Pearson residuals obtained after the application of the linear model (1) do not deviate from a white noise sequence. The AIC (BIC, respectively) of the fit is equal to -7273.609 (-7279.609, respectively).

To apply the non-linear model (5), we first choose the value of the parameter γ by following a profiling procedure whereby we calculate the log-likelihood function (10) for a grid of values of γ and then we choose γ as the value that maximizes the log-likelihood function—Fig. 4 illustrates the method. With $\gamma = 1.80$, we fit model (5) to the mortality data and we obtain that $\hat{d} = 5.072$ (0.007), $\hat{a} = 0.886$ (0.194) and $\hat{b} = 0.089$ (0.212), where in parentheses are the standard errors of the estimators obtained by inversion of the information matrix, see Theorem 1. Note that both the constant term and the coefficient that corresponds to feedback mechanism are significant, whereas the Y_{t-1}-terms is less important in explaining the dynamics of the observed process. Nevertheless, the sum of both coefficients is close to unity which

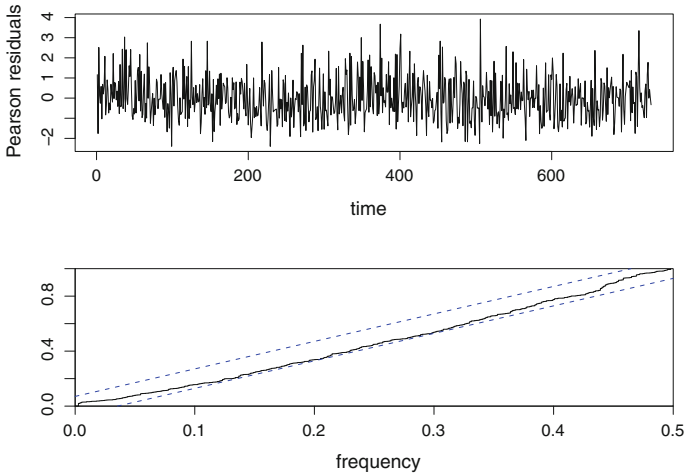

Fig. 3 Pearson residuals (*top*) and their cumulative periodogram plot (*bottom*) for the linear model (1) fitted to mortality data

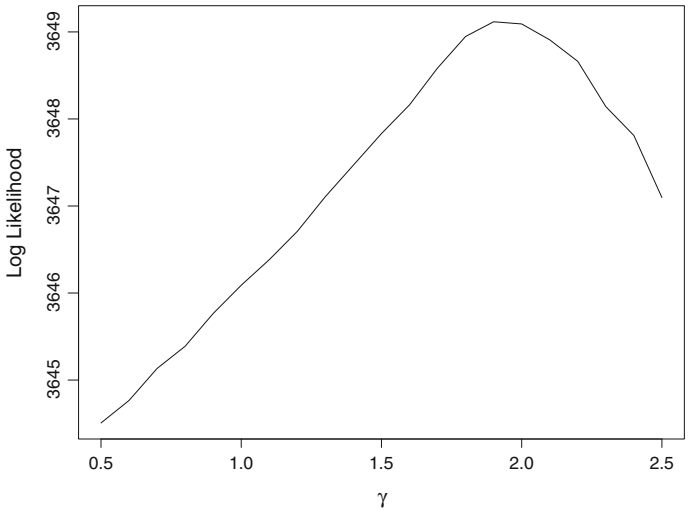

Fig. 4 Log-likelihood function (10) for a grid of values of γ. The values of γ are generated from 0.50 to 2.50 by a step equal to 0.10. For each value of γ, model (5) is fitted to the mortality data and the corresponding log-likelihood is evaluated. For these data, the maximum value of γ occurs at 1.80

possibly indicates some evidence of non-stationarity when using model (5) to explain these data. In contrast, the sum of the estimated coefficient for the linear model fit is 0.874 which is well below the unity. Some further analysis of model (5) shows that the mean square error of the Pearson residuals is 1.130 which is slightly smaller than the corresponding mean square error of the Pearson residuals obtained from the linear model fit. The AIC (BIC, respectively) for model (5) fit is −7291.895 (−7297.895, respectively). We see that both of these values are smaller than the corresponding

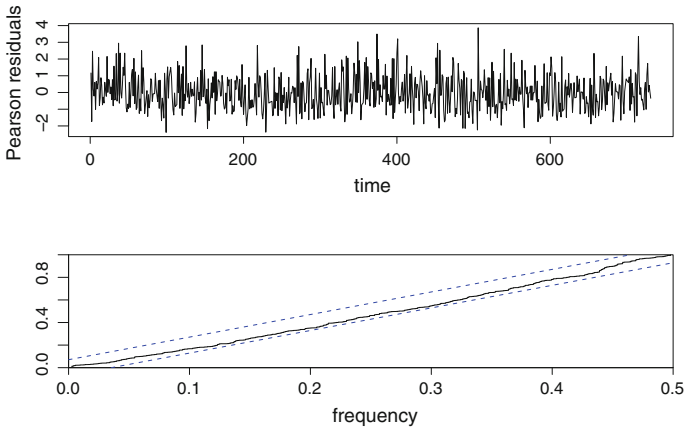

Fig. 5 Pearson residuals (*top*) and their cumulative periodogram plot (*bottom*) for the non-linear model (5) fitted to mortality data

values obtained form the linear model. As a final remark, we note that Fig. 5 points to the adequacy of the model when applied to these mortality data. In fact, the cumulative periodogram plot shown in the bottom plot of Fig. 5 shows that the residuals obtained from the fit of model (5) approximate better a white noise processes than the corresponding residuals obtained form the linear model fit; compare with bottom plot of Fig. 3.

Appendix

Proof of Lemma 1

This lemma is concerned with the convergence of $\mathbf{G}^m$ to $\mathbf{G}$ and the positive definiteness of these two matrices. It is based on the evaluation of derivatives of λ_t and λ_t^m with respect to the unknown parameters. For simplicity, in this proof and in other proofs, we only treat derivatives with respect to $\boldsymbol{\theta}_1$; the derivation for $\boldsymbol{\theta}_2$ being quite similar.

In the present case we have

$$E\left\|\frac{\partial\lambda_t^m}{\partial\boldsymbol{\theta}_1}-\frac{\partial\lambda_t}{\partial\boldsymbol{\theta}_1}\right\| = E\left\|\frac{\partial f(\lambda_{t-1}^m,\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}-\frac{\partial f(\lambda_{t-1}^m,\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1} + \frac{\partial f(\lambda_{t-1}^m,\boldsymbol{\theta}_1)}{\partial\lambda}\frac{\partial\lambda_{t-1}^m}{\partial\boldsymbol{\theta}_1}-\frac{\partial f(\lambda_{t-1},\boldsymbol{\theta}_1)}{\partial\lambda}\frac{\partial\lambda_{t-1}}{\partial\boldsymbol{\theta}_1}\right\|.$$

But by using an addition subtraction argument, the last two terms can be written:

$$\frac{\partial f(\lambda_{t-1}^m,\boldsymbol{\theta}_1)}{\partial\lambda}\left[\frac{\partial\lambda_{t-1}^m}{\partial\boldsymbol{\theta}_1}-\frac{\partial\lambda_{t-1}}{\partial\boldsymbol{\theta}_1}\right]+\left[\frac{\partial f(\lambda_{t-1}^m,\boldsymbol{\theta}_1)}{\partial\lambda}-\frac{\partial f(\lambda_{t-1},\boldsymbol{\theta}_1)}{\partial\lambda}\right]\frac{\partial\lambda_{t-1}}{\partial\boldsymbol{\theta}_1}. \quad (16)$$

By (13), (14), assumptions A1, A4 and the existence of any moment of λ_t and λ_t^m, using the technique of proof of Proposition 3, it follows that $E||\partial\lambda_{t-1}/\partial\boldsymbol{\theta}_1||^2 < \infty$ and $E||\partial\lambda_{t-1}^m/\partial\boldsymbol{\theta}_1||^2 < \infty$. Then, by the Schwartz inequality, assumption A4, Proposition 3 and Lebesgue dominated convergence, the expected value of the norm of the last of the two terms in (16) can be made arbitrarily small. We are then left with a recursive relationship of the type (9) in the proof of Proposition 3 and using A1,

$$E\left\|\frac{\partial\lambda_t^m}{\partial\boldsymbol{\theta}_1}-\frac{\partial\lambda_t}{\partial\boldsymbol{\theta}_1}\right\|<\delta_m,$$

with $\delta_m \to 0$ as $m \to \infty$. From this it follows that the analogue of Fokianos et al. (2009b, Eq. A–4) in their proof of Lemma 3.1 is fulfilled.

We can handle differentiation with respect to $\boldsymbol{\theta}_2$ with very similar arguments. Higher order moments like $E(||\partial\lambda_t^m/\partial\boldsymbol{\theta}_1 - \partial\lambda_t/\partial\boldsymbol{\theta}_1||^2)$ can also be tackled using such arguments. Finally, there is the following term

$$E\left\|\frac{1}{\lambda_t^m}\frac{\partial\lambda_t^m}{\partial\boldsymbol{\theta}_1}\frac{\partial\lambda_t^m}{\partial\boldsymbol{\theta}_1'}-\frac{1}{\lambda_t}\frac{\partial\lambda_t}{\partial\boldsymbol{\theta}_1}\frac{\partial\lambda_t}{\partial\boldsymbol{\theta}_1'}\right\|,$$

analogous to the term treated after the proof of Fokianos et al. (2009b, Eq. A–4). In the linear case we used $\lambda_t \geq d > 0$ to evaluate that term, whereas in the present case we use the defining relationships (3) and (4) and assumption A2. Finally, the positive definiteness of the information matrix follows from the convergence established in the first part of the lemma and assumption A3.

Proof of Lemma 2

This lemma is concerned with the asymptotic normality of $\widehat{\boldsymbol{\theta}}^m$ and the convergence to this distribution as $m \to \infty$. This is achieved by using the property of geometric ergodicity for the perturbed system and by evaluating the appropriate derivatives and their differences. To apply the martingale CLT we need to verify a Lindeberg condition which means that we essentially have to prove boundedness of moments such as

$$E\left\|\frac{\partial\lambda_t^m}{\partial\boldsymbol{\theta}_1}\right\|^4<\infty,\quad E\left\|\frac{\partial\lambda_t^m}{\partial\boldsymbol{\theta}_2}\right\|^4<\infty,\quad E\left(\frac{Y_t^m}{\lambda_t^m}\right)^4<\infty. \tag{17}$$

From the model definition (4) and Assumption A2 we obtain that $\lambda_t^m \geq f(\lambda_{t-1}^m) \geq c_1$. The last inequality of (17) then follows and by using the existence of arbitrary moments of Y_t^m and the fact that $1/\lambda_t^m \leq 1/c_1$. Next, observe that by repeated substitution and use of assumption A1

$$\begin{aligned}\left\|\frac{\partial\lambda_t^m}{\partial\boldsymbol{\theta}_1}\right\| &= \left\|\frac{\partial f(\lambda_{t-1}^m,\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}+\frac{\partial f(\lambda_{t-1}^m,\boldsymbol{\theta}_1)}{\partial\lambda}\frac{\partial\lambda_{t-1}^m}{\partial\boldsymbol{\theta}_1}\right\| \\ &\leq \sum_i \alpha_2^{i-1}\left\|\frac{\partial f(\lambda_{t-i}^m,\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\|+\alpha_2^t\left\|\frac{\partial\lambda_0^m}{\partial\boldsymbol{\theta}_1}\right\|,\end{aligned} \tag{18}$$

where $\alpha_2 < 1$ is given in assumption A1. The last term can be neglected by an appropriate choice of initial condition. On the other hand

$$E\left\|\sum_i \alpha_2^{i-1}\frac{\partial f(\lambda_{t-i}^m,\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\| \le \sum_i \alpha_2^{i-1}E|\lambda_{t-i}^m| < \infty, \tag{19}$$

due to assumption A4 and the existence of the first moment of $|\lambda_t^m|$. We use this expansion raised to the fourth power to bound $E||\partial\lambda_t^m/\partial\boldsymbol{\theta}_1||^4$. We will then encounter moments of the type $E|\lambda_{t-i_1}^m\lambda_{t-i_2}^m\lambda_{t-i_3}^m\lambda_{t-i_4}^m|$ which exist due to the Schwartz inequality and the existence of an arbitrary moment of λ_t. Since $\alpha_2 < 1$ the corresponding four-dimensional sum is finite, independent of t, and the first part of (17) is proved. The rest of the proof follows step by step from the proof of Fokianos et al. (2009b, Lemma 3.2).

Proof of Lemma 3

This lemma is concerned with the convergence of the scaled Hessian $n^{-1}\mathbf{H}_n^m$ to the matrix $\mathbf{G}^m$ corresponding to $\mathbf{G}$ of Theorem 1 as $n \to \infty$, the convergence of $\mathbf{G}^m$ to $\mathbf{G}$ as $m \to \infty$ and the convergence of $\mathbf{H}_n^m - \mathbf{H}_n$ to zero as $n, m \to \infty$.

To apply the LLN of Jensen and Rahbek (2007), we have to show that

$$E\left\|\frac{\partial^2\lambda_t^m}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\|^2 < \infty \quad \text{and} \quad E(Z_t^m)^2 < \infty, \tag{20}$$

where $Z_t^m = (Y_t^m/\lambda_t^m - 1)$. Moreover, to carry through the last part of the proof one has to show that

$$\left\|\frac{\partial^2\lambda_t^m}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2\lambda_t}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\| \xrightarrow{\text{a.s}} 0, \tag{21}$$

as $m \to \infty$. The second part of (20) follows immediately from $E((Z_t^m)^2|\mathcal{F}_{t-1}) = 1/\lambda_t^m \le 1/c_1$. The first part of (20) is more complicated and requires the evaluation of mixed second derivatives. As an example consider

$$\frac{\partial^2\lambda_t^m}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1'} = \frac{\partial^2 f}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1'} + \frac{\partial^2 f}{\partial\lambda\partial\boldsymbol{\theta}_1}\frac{\partial\lambda_{t-1}^m}{\partial\boldsymbol{\theta}_1'} + \frac{\partial^2 f}{\partial\lambda^2}\frac{\partial\lambda_{t-1}^m}{\partial\boldsymbol{\theta}_1}\frac{\partial\lambda_{t-1}^m}{\partial\boldsymbol{\theta}_1'} + \frac{\partial f}{\partial\lambda}\frac{\partial^2\lambda_{t-1}^m}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1'}, \tag{22}$$

where the argument of $f(\cdot)$ is always $(\lambda_{t-1}^m, \boldsymbol{\theta}_1)$ We can now use the same expansion as in (18) and (19) in the modified proof of Fokianos et al. (2009b, Lemma 3.2). This is combined with the Schwarz inequality, assumptions A4 (in particular the last statement of this condition), and the existence of any moments of Y_t^m and λ_t^m yield the desired result in (20). To prove (21), the representation (22) is employed again. The differences arising from the three last terms can be handled using the technique of the last part of the proof of Fokianos et al. (2009b, Lemma 3.1), whereas the difference

due to the first term of (22) can be treated by a first-order Taylor expansion in λ, use of A4 and Proposition 3.

Proof of Lemma 4

This lemma is concerned with the evaluation of the third-order derivative of the log likelihood. If $O(\boldsymbol{\theta}_0)$ is a fixed open neighborhood of $\boldsymbol{\theta}_0$, then we must show that

$$m_t = \sup_{\boldsymbol{\theta} \in O(\boldsymbol{\theta}_0)} \frac{\partial^3 l_t(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_k}$$

should exist and be bounded as $\boldsymbol{\theta}$ varies. Moreover, it has to be shown that

$$E(|m_t|) < \infty \quad \text{and} \quad |m_t^m - m_t| \overset{\text{a.s}}{\to} 0,$$

as $m \to \infty$. The first claim is shown using the existence of at least fourth-order continuous derivatives. Along these lines, it is sufficient to show that for $M_t = M_{t;i,j,k}(\boldsymbol{\theta}) = \partial^3 l_t / \partial\theta_i \partial\theta_j \partial\theta_k$,

$$E(|M_t|) < \infty \quad \text{and} \quad |M_t^m - M_t| \overset{\text{a.s}}{\to} 0.$$

The verification of the above display requires some detailed calculations which are based on the fact that three different types of terms are involved,

$$\left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1\right) \frac{\partial^3 \lambda_t(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_k}, \quad \frac{Y_t}{\lambda_t^2(\boldsymbol{\theta})} \frac{\partial\lambda_t(\boldsymbol{\theta})}{\partial\theta_i} \frac{\partial^2 \lambda_t(\boldsymbol{\theta})}{\partial\theta_j \theta_k}, \quad \frac{Y_t}{\lambda_t^3(\boldsymbol{\theta})} \frac{\partial\lambda_t(\boldsymbol{\theta})}{\partial\theta_i} \frac{\partial\lambda_t(\boldsymbol{\theta})}{\partial\theta_j} \frac{\partial\lambda_t(\boldsymbol{\theta})}{\partial\theta_k}.$$

These can now be evaluated using the technique of the proofs of Fokianos et al. (2009b, Lemmas 3.1 and 3.2) and again using the assumptions of A1–A4. The condition $|M_t^m - M_t| < \delta_m \to 0$ is shown using arguments that are identical to those used at the end of the proof of Lemma 3.

Proof of Theorem 1

Recall the log-likelihood function (10) of the unperturbed model. Let $C_n(r) = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| \leq r/\sqrt{n}\}$ be a compact neighborhood of the true value $\boldsymbol{\theta}_0$ for any $r > 0$. Then, if $\boldsymbol{\theta}^\star$ lies on the line between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, a Taylor expansion, shows that

$$\begin{aligned} l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0) &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{S}_n(\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{H}_n(\boldsymbol{\theta}^\star)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' (\mathbf{S}_n(\boldsymbol{\theta}_0) - \mathbf{S}_n^m(\boldsymbol{\theta}_0)) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{S}_n^m(\boldsymbol{\theta}_0) \\ &\quad - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' (\mathbf{H}_n(\boldsymbol{\theta}^\star) - \mathbf{H}_n^m(\boldsymbol{\theta}^\star))(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'(\mathbf{H}_n^m(\boldsymbol{\theta}^\star)-\mathbf{H}_n^m(\boldsymbol{\theta}_0))(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\
&-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{H}_n^m(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\
&= I_{1n}+I_{2n}+I_{3n}+I_{4n}+I_{5n}. \qquad (23)
\end{aligned}
$$

We will prove that for every $\eta > 0$, there exist n and r such that

$$
\mathrm{P}[l(\boldsymbol{\theta})-l(\boldsymbol{\theta}_0) < 0,\ \forall\,\boldsymbol{\theta}\in\partial C_n(r)] \geq 1-\eta,
$$

which shows that the maximum is attained in the interior of $C_n(r)$, with probability tending to 1. However, we have that $I_{in} \to 0$, in probability, for $i = 1, 3, 4$. To see this, note that the result for I_{1n} is based on the proof of Lemma 2. The other two convergence results are based on the proofs of Lemmas 3 and 4 and the continuity of the Hessian matrix as a function of $\boldsymbol{\theta}$. In addition, we obtain that $I_{2n} \leq ||\mathbf{S}_n^m(\boldsymbol{\theta}_0)||r/\sqrt{n}$ and $I_{5n} \leq -\lambda_{\min}(\mathbf{H}_n^m(\boldsymbol{\theta}_0))r^2/2n$, where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a symmetric matrix A. Therefore, combining all the above, (23) yields

$$
\begin{aligned}
\mathrm{P}[l(\boldsymbol{\theta})-l(\boldsymbol{\theta}_0) < 0,\ \forall\,\boldsymbol{\theta}\in\partial C_n(r)] &\geq \mathrm{P}\left[\left\|\frac{\mathbf{S}_n^m(\boldsymbol{\theta}_0)}{\sqrt{n}}\right\|^2 \leq \frac{1}{4}r^2\lambda_{\min}^2(\mathbf{H}_n^m(\boldsymbol{\theta}_0)/n)\right]\\
&\geq \mathrm{P}\left[\left\|\frac{\mathbf{S}_n^m(\boldsymbol{\theta}_0)}{\sqrt{n}}\right\|^2 \leq \frac{1}{4}r^2\right]-\frac{\eta}{2}\\
&\geq 1-\frac{4E[||\mathbf{S}_n^m(\boldsymbol{\theta}_0)/\sqrt{n}||^2]}{r^2}-\frac{\eta}{2}.
\end{aligned}
$$

Since $E[||\mathbf{S}_n^m(\boldsymbol{\theta}_0)/\sqrt{n}||^2] < \infty$ from Lemma 2, the second term of the above expression can become arbitrarily small. Therefore, we have established asymptotic existence, or in other words, there exists a sequence of conditional MLE, $\widehat{\boldsymbol{\theta}}$ such that, for any $\eta > 0$, there exists an r and an n_1 such that

$$
P[\widehat{\boldsymbol{\theta}} \in C_n(r)] \geq 1-\eta, \quad \text{for all } n \geq n_1.
$$

Similar to the proofs of Lemmas 2 and 3. we also obtain that $\mathbf{H}_n(\boldsymbol{\theta})$ is positive definite throughout $C_n(r)$ with probability tending to one. In conclusion, there is exactly one solution $\widehat{\boldsymbol{\theta}}$ to the likelihood equation in the interior of $C_n(r)$.

Using the same argument as before, for any $\delta, 0 < \delta < r$, there exists with probability tending to one a solution to the score equations in $C_n(\delta)$. But $\widehat{\boldsymbol{\theta}}$ is the unique solution to the score equations in $C_n(r)$ and therefore lies in $C_n(\delta)$ with probability tending to one. In other words, $\widehat{\boldsymbol{\theta}}$ is consistent. The asymptotic normality follows by a Taylor expansion of the score and Lemmas 1–4.

Acknowledgments We would like to thank the Associate Editor and two reviewers for helpful comments and suggestions. D. Gomes provided the mortality data. Research was supported by Cyprus Research Foundation, Grant PROSELKISI/PROEM/0308/01.

References

Berkes, I., Horváth, L., Kokoszka, P. (2003). GARCH processes: structure and estimation. *Bernoulli 9*, 201–227.

Brockwell, P. J., Davis, R. A. (1991). *Time series: data analysis and theory* (2nd ed.). New York: Springer.

Brumback, B. A., Ryan, L. M., Schwartz, J. D., Neas, L. M., Stark, P. C., Burge, H. A. (2000). Transitional regression models with application to environmental time series. *Journal of the American Statistical Association 85*, 16–27.

Chen, C. W. S., Gerlach, R., Choy, B., Lin, C. (2010). Estimation and inference for exponential smooth transition nonlinear volatility models. *Journal of Statistical Planning and Inference 140*, 719–733.

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

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

Doukhan, P., Fokianos, K., Tjøstheim, D. (2012). On weak dependence conditions for Poisson autoregressions *Statistics & Probability Letters* (to appear).

Fahrmeir, L., Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer.

Fan, J., Yao, Q. (2003). *Nonlinear time series*. New York: Springer-Verlag.

Fokianos, K., Kedem, B. (2004). Partial likelihood inference for time series following generalized linear models. *Journal of Time Series Analysis 25*, 173–197.

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

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

Fokianos, K., Rahbek A., Tjøstheim, D. (2009b). Poisson autoregression (complete version). http://pubs.amstat.org/toc/jasa/104/488.

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

Franke, J. (2010). Weak dependence of functional INGARCH processes (unpublished manuscript).

Gao, J., King, M., Lu, Z., Tjøstheim, D. (2009). Specification testing in nonlinear and nonstationary time series regression. *Annals of Statistics 37*, 3893–3928.

Haggan, V., Ozaki, T. (1981). Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika 68*, 189–196.

Jensen, S. T., Rahbek, A. (2007). On the law of large numbers for (geometrically) ergodic Markov chains. *Econometric Theory 23*, 761–766.

Jung, R. C., Kukuk, M., Liesenfeld, R. (2006). Time series of count data: modeling, estimation and diagnostics. *Computational Statistics & Data Analysis 51*, 2350–2364.

Kedem, B., Fokianos, K. (2002). *Regression models for time series analysis*. Hoboken, NJ: Wiley.

Li, W. K. (1994). Time series models based on generalized linear models: some further results. *Biometrics 50*, 506–511.

MacDonald, I. L., Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.

Meitz, M., Saikonnen, P. (2008). Ergodicity, mixing and existence of moments of a class of Markov models with applications to GARCH and ACD models. *Econometric Theory 24*, 1291–1320.

Meyn, S. P., Tweedie, R. L. (1993). *Markov Chains and stochastic stability*. London: Springer.

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

Teräsvirta, T., Tjøstheim, T., Granger, C. W. J. (2010). *Modelling nonlinear economic time series*. Oxford: Oxford University Press.

Tong, H. (1990). *Nonlinear time series: a dynamical system approach*. New York: Oxford University Press.

Zeger, S. L., Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics 44*, 1019–1031.