Discussion Papers
Department of Economics
University of Copenhagen

No. 08-35

Poisson Autoregression

Konstantinos Fokianos, Anders Rahbek,
and Dag Tjøstheim

Studiestræde 6, DK-1455 Copenhagen K., Denmark
Tel.: +45 35 32 30 82 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (online)

Poisson Autoregression

Konstantinos Fokianos[1] Anders Rahbek[2] Dag Tjøstheim[3]

[1]Department of Mathematics & Statistics, University of Cyprus
e-mail: fokianos@ucy.ac.cy

[2]Department of Economics, University of Copenhagen and CREATES
e-mail:rahbek@econ.ku.dk

[3]Department of Mathematics, University of Bergen
e-mail:Dag.Tjostheim@math.uib.no

Abstract

This paper considers geometric ergodicity and likelihood based inference for linear and nonlinear Poisson autoregressions. In the linear case the conditional mean is linked linearly to its past values as well as the observed values of the Poisson process. This also applies to the conditional variance, implying an interpretation as an integer valued GARCH process. In a nonlinear conditional Poisson model, the conditional mean is a nonlinear function of its past values and a nonlinear function of past observations. As a particular example an exponential

autoregressive Poisson model for time series is considered. Under geometric ergodicity the maximum likelihood estimators of the parameters are shown to be asymptotically Gaussian in the linear model. In addition we provide a consistent estimator of the asymptotic covariance, which is used in the simulations and the analysis of some transaction data. Our approach to verifying geometric ergodicity proceeds via Markov theory and irreducibility. Finding transparent conditions for proving ergodicity turns out to be a delicate problem in the original model formulation. This problem is circumvented by allowing a perturbation of the model. We show that as the perturbations can be chosen to be arbitrarily small, the differences between the perturbed and non-perturbed versions vanish as far as the asymptotic distribution of the parameter estimates is concerned.

Keywords: generalized linear models, non-canonical link function, count data, Poisson regression, likelihood, geometric ergodicity, integer GARCH, ϕ-irreducibility, observation driven models, asymptotic theory.

1 Introduction

In this paper we study ergodicity and likelihood inference for a specific class of GARCH type Poisson time series models. The necessity for such an investigation arises from the fact that count dependent sequences appear in several diverse scientific fields, including medical, environmental or financial applications. As an illustrative example, consider the upper plot of Figure 2 which depicts the number of transactions per minute for some stock during a specific day. These data consist of a count time series and therefore statistical methodology should be developed for time series modeling, estimation, inference and prediction.

Models for time series of counts have been considered by many authors–for a comprehensive account see Kedem and Fokianos (2002, Ch. 4), for instance. The most popular choice among several authors is the log–linear model. In other words, if it is assumed that Y_t is conditionally Poisson distributed with mean λ_t, then most existing models are based upon regressing $\log \lambda_t$–the canonical link parameter–on past values of the response and/or covariates. As it has been developed by Fokianos and Kedem (2004), these models fall within the broad class of generalized linear

time series models, and their analysis is based on partial likelihood inference. Estimation, diagnostics, model assessment, and forecasting are implemented in a straightforward manner where the computation is carried out by a number of the existing statistical computing environments. In addition, empirical evidence shows that both positive and negative association can be taken into account by a suitable parametrization of the model, see e.g. Zeger and Qaqish (1988).

A largely missing element in these developments has been the possibility of an autoregressive feedback mechanism in $\{\lambda_t\}$. Such a feedback is a key feature in state space models. One example is the GARCH model for volatility. These models are generally expected to be more parsimonious. The purpose of this paper is to study autoregressive models of λ_t–not $\log \lambda_t$–both linear and non-linear. More specifically, λ_t is regressed on past values of the observed process and past values of λ_t itself–this model has been considered by Rydberg and Shephard (2000), Streett (2000) and more recently by Ferland et al. (2006). Two classes of models are actually proposed in the present paper. The first class is a simple linear model and it is given by expression (1) in Section 2.1. This model states that the conditional mean of the Poisson observed time series is a linear function of its past values and lagged values of the observations. The model can be motivated by the arguments of Rydberg and Shephard (2000) who show that it is a reasonable approximation for inference about the number of trades within a small time interval. Furthermore, model (1) is similar to the GARCH models (Bollerslev (1986)) in the sense that for the Poisson distribution the conditional mean equals the conditional variance. In fact, this is a generalized linear model for time series of counts but with an identity link–that is a non canonical link function.

The second class of models generalizes (1) by imposing a non linear structure on both past values of λ_t and lagged values of Y_t. Accordingly this is a non-linear model for Poisson time series and can be cast as a generalized non-linear model for time series of counts, see equation (4) in Section 2.3. A specific example which falls within this framework is the Poisson analogue of the so called exponential autoregressive model, see Haggan and Ozaki (1981). In the Poisson time series framework, this model is formulated in (5).

The contribution of this article is to study ergodicity and likelihood inference for both types of models. A fundamental complication with the analysis of these models is the proof of the geo-

metric ergodicity of both the observed $\{Y_t\}$ and latent process $\{\lambda_t\}$. This problem is bypassed by showing that the slightly perturbed models (3) and (6) are geometrically ergodic under simple restrictions on the parameter space. The perturbation idea is a form of regularization–an idea similar in spirit to the analysis of ill–posed problems. Likelihood inference is developed in detail for the linear model by showing that the difference between the perturbed and unperturbed model can be made arbitrarily small provided that the perturbation decreases. It is shown that the corresponding maximum likelihood estimator is asymptotically normal. The results are applied to real and simulated data and several guidelines for the development of algorithms for numerical maximization of the log likelihood for both type of models are given. Non linear models for count time series have not been discussed in the literature and their development adds an additional tool to the analysis of count time series.

The organization of the presentation is as follows. Section 2 introduces the reader to the relevant models and discusses the link between their perturbed and unperturbed versions. In addition, Propositions 2.1–2.4 show the joint ergodicity of both the observed and unobserved process for linear and non linear perturbed models. Section 3 develops likelihood inference for the linear model (3) and discusses some aspects of inference for the non linear model (6) with a special focus on the exponential autoregressive model (7). Several simulations and some real data analysis are presented in Section 3 while the presentation concludes with an appendix which contains the proofs of some of the theoretical results. A complete version of this article with all proofs included is located at `web address`.

2 Model Specification & Ergodicity Results

Suppose that $\{Y_t\}$ is a time series of counts and assume that $\mathcal{F}_t^{Y,\lambda}$ stands for the σ–field generated by $\{Y_0, \ldots, Y_t, \lambda_0\}$, that is $\mathcal{F}_t^{Y,\lambda} = \sigma(Y_s, s \leqslant t, \lambda_0)$, where $\{\lambda_t\}$ is a Poisson intensity process to be introduced below. Initially, a linear Poisson GARCH type model is considered and then a non linear model is proposed.

2.1 Linear Model

Consider the following model given by,

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

for $t \geqslant 1$ and where the parameters d, a, b are assumed to be positive. In addition assume that λ_0 and Y_0 are fixed. Recall that for the Poisson distribution, the conditional mean is equal to the conditional variance, that is $\text{E}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \text{Var}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \lambda_t$. Therefore it is tempting to call (1) an INGARCH(1,1)–that is integer GARCH–model since its structure parallels that of the customary GARCH model, see Bollerslev (1986). However, although such a definition is quite plausible for this particular case, the proposed modeling is based on the evolution of the mean of the Poisson instead of its variance. In other words, model (1) specifies a conditional mean relation to the past values of both λ_t and Y_t.

For technical reasons having to do with the proofs of asymptotic normality of parameter estimates, it turns out that it is advantageous to rephrase the model formulation (1). More specifically, it is desirable to express the sequence of independent Poisson drawings–that is the first equation of (1)–more explicitly in terms of random variables, like the observational equation in a state space model, or like the defining equation in a GARCH model giving the relationship between the observations and the conditional variance. To achieve this, for each time point t, introduce a Poisson process $N_t(\cdot)$ of unit intensity. Then, the first equation of (1) can be restated in terms of these Poisson processes by assuming that Y_t is equal to the number of events $N_t(\lambda_t)$ 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 process of unit intensity and rephrase (1) as

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

for $t \geqslant 1$ and with Y_0, λ_0 fixed. This notation will be used throughout the paper, and it is emphasized that we can always recover (1) by (2).

Model (1) (or its rephrasing (2)) is related to the theory of generalized linear models for time series, see Kedem and Fokianos (2002, Ch. 1 & 4). In particular, the driving random component

of the model corresponds to the Poisson distribution which belongs to the exponential family of distributions. The link function is taken to be the identity while the systematic component is the time dependent random vector $(1, \lambda_{t-1}, Y_{t-1})'$. Hence, (2) is a non canonical link model for time series of counts. Furthermore, notice that even though the vector of time dependent covariates that influences the evolution of (1) is composed by the unobserved process λ_t, the linear model still belongs to the class of observation driven models as defined by Cox (1981). This is so because the unobserved process λ_t can be expressed as a function of past values of the observed process Y_t, after repeated substitution. Observation driven models for time series of counts have been studied by several authors including Zeger and Qaqish (1988), Li (1994) and more recently by Brumback et al. (2000), Fahrmeir and Tutz (2001), Benjamin et al. (2003), Davis et al. (2003) and Jung et al. (2006). However, a log–linear model for the mean of the observed process is usually assumed and its structure is composed by past values of the response, moving average terms and other explanatory variables. With the exception of Davis et al. (2003), who considers a simple but important case of a log-linear model, all the other references do not discuss the problem of ergodicity of the joint process (Y_t, λ_t) which is instrumental in developing asymptotic estimation theory.

Second order properties of model (1) have been studied by Rydberg and Shephard (2000) while Streett (2000) shows existence and uniqueness of a stationary distribution. In a recent contribution, Ferland et al. (2006) considered model (1) in a more general form and it was shown that the process Y_t is stationary provided that $0 \leqslant a+b < 1$, by employing a different technique from that of Streett (2000). In particular, $\mathrm{E}[Y_t] = \mathrm{E}[\lambda_t] \equiv \mu = d/(1-a-b)$ and its autocovariance function is given by

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

In addition, it was shown that all moments of model (1) are finite if and only if $0 \leqslant a+b < 1$.

Upon noticing that

$$\text{Var}[Y_t] = \mu \left(1 + \frac{b^2}{1-(a+b)^2} \right),$$

we conclude that $\text{Var}[Y_t] \geqslant \text{E}[Y_t]$ with equality when $b = 0$. Thus, the inclusion of the past values of Y_t in the evolution of λ_t leads to overdispersion–a phenomenon that occurs frequently in count time series data. In what follows, we discuss some properties of the linear model and show ergodicity for a modification of (2).

2.2 Ergodicity of a perturbed model

We first note that $\{\lambda_t\}$ defined by (2) is a Markov chain. Consider the skeleton $\lambda_t = d + a\lambda_{t-1}$ of (2). Then $\lambda^* = d/(1-a)$ is the solution of $\lambda = d+a\lambda$, i.e. a fix point of the mapping $f(\lambda) = d+a\lambda$. In Streett (2000) it was shown that if $0 \leqslant a + b < 1$, there exists a stationary initial distribution for $\{\lambda_t\}$ by showing that the point λ^* is reachable, where (cf. Meyn and Tweedie (1993, p.131)) a point λ^1 in the state space Λ is reachable if for every neighborhood O of λ^1, $\sum_n \text{P}^n(\lambda, O) > 0$, $\lambda \in \Lambda$. Here $\text{P}^n(\lambda, A) = \text{P}(\lambda_n \in A \mid \lambda_0 = \lambda)$, is the n'th step transition probability of $\{\lambda_t, t \geqslant 0\}$.

In the course of the proof of Lemma A-1 in the appendix, it will be shown that every point in $[\lambda^*, \infty)$ is reachable, if $0 < a < 1$, i.e. $\{\lambda_t\}$ is open set irreducible on $[\lambda^*, \infty)$. However, in our approach to show geometric ergodicity we need ϕ–irreducibility, where ϕ is the Lebesgue measure with support $[k, \infty)$, for some $k \geqslant \lambda^*$. Establishing ϕ–irreducibility is equivalent to showing that $\sum_n \text{P}^n(\lambda, A) > 0$ if $\phi(A) > 0$. The problem is that there are sets A of positive Lebesgue measure that are not open sets, for example the set of all irrationals in $[\lambda^*, \infty)$, and Lemma A-1 cannot be used directly to handle such sets. One could add an assumption stating that the chain is not allowed to stay on sets of Lebesgue measure zero, but since it is not clear how restrictive such an assumption is, we have decided to avoid that issue by resorting to the perturbed chain (Y_t^m, λ_t^m) defined by

$$Y_t^m = N_t\left(\lambda_t^m\right), \ \lambda_t^m = d + a\lambda_{t-1}^m + bY_{t-1}^m + \varepsilon_{t,m}, \tag{3}$$

with λ_0^m, Y_0^m fixed, and

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

where $1(\cdot)$ is the indicator function, and where $\{U_t\}$ is a sequence of iid uniform random variables on $(0,1)$ and such that the $\{U_t\}$ is independent of $\{N_t(\cdot)\}$. Note that the introduction of $\{U_t\}$ makes it possible to establish irreducibility, see the proofs of Lemma A-1 and Proposition 2.1 Another possibility would be to try an approach to prove ergodicity that does not depend on ϕ-irreducibility, see, e.g Aue et al. (2006), Mikosch and Straumann (2006). But the structure of our model is different, and we have not succeeded in using such an approach. Note that $\{\lambda_t^m\}$ is still a Markov chain. Our strategy is to prove geometric ergodicity of $\{Y_t^m, \lambda_t^m\}$ defined by (3). Then to use this to obtain asymptotic normality for the likelihood estimators of (3), and finally by letting $c_m \to 0$ obtain asymptotic normality of the likelihood estimates of (2).

The perturbation in (3) is a purely auxiliary device to obtain ϕ–irreducibility. The U_t's could be thought of as pseudo observations generated by the uniform law. As will become clear in the subsequent proofs, the perturbation can be introduced in many other ways. For, instance, it is enough to set $\{U_t\}$ to be an i.i.d sequence of positive random variables with bounded support possessing density on the positive real axis with respect to the Lebesgue measure and finite fourth moment. In addition, the form of the likelihood functions for $\{Y_t\}$ and $\{Y_t^m\}$ as far as dependence on $\{\lambda_t\}$ is concerned will be the same for both models (2) and (3). Note that both $\{Y_t\}$ and $\{Y_t^m\}$ can be identified with the observations in the likelihood, but they cannot be identified as stochastic variables since they are generated by different models. Technically speaking, with the introduction of $\{\varepsilon_{t,m}\}$, the process $\{\lambda_t^m\}$ is made into a T-chain with a continuous component, where open set irreducibility implies measure theoretic irreducibility.

We have the following results concerning ergodicity of model (3). Their proof is postponed to the Appendix. There it is first proved that the unobserved process $\{\lambda_t^m\}$ is geometrically ergodic.

Proposition 2.1 Consider model (3) and suppose that $0 < a+b < 1$. Then the process $\{\lambda_t^m, t \geqslant 0\}$ is a geometrically ergodic Markov chain with finite moments of order k, for an arbitrary k.

The above proposition is useful in obtaining geometric ergodicity of the joint process $(Y_t^m, U_t, \lambda_t^m)$. In fact, it can been shown that the joint process is $V_{(Y,U,\lambda)}$–geometrically ergodic with $V_{Y,U,\lambda}(Y,U,\lambda) = 1 + Y^k + \lambda^k + U^k$, for a definition see Meyn and Tweedie (1993, p.355). In particular, Proposition 2.2 shows the existence of moments of $(Y_t^m, U_t, \lambda_t^m)$, for any k. In addition, the ergodicity result is employed in the next section where the large sample estimation theory is studied. Note that inferential results for the $\{Y_t^m\}$ process depend upon proving that geometric ergodicity of the $\{\lambda_t^m\}$ series implies geometric ergodicity of the chain $\{(Y_t^m, U_t, \lambda_t^m)\}$. The proof of the following proposition follows closely the arguments of Meitz and Saikonnen (2008) (see also Carrasco and Chen (2002)) and its proof is indicated in the appendix.

Proposition 2.2 Consider model (3) and suppose that the conditions of Proposition 2.1 hold. Then the process $\{(Y_t^m, \lambda_t^m, U_t), t \geqslant 0\}$ is a $V_{(Y,U,\lambda)}$–geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y,U,\lambda) = 1 + Y^k + \lambda^k + U^k$.

The following lemma quantifies the difference between (2) and (3), as $m \to \infty$ such that $c_m \to 0$, and shows that essentially the perturbed model can be made arbitrarily close to the unperturbed model. The crucial condition is that the sum of a and b must be less than one –this as seen from Proposition 2.1 was also the natural condition for proving ergodicity of (3). It is in the proof of this lemma that the rephrasing of model (1) as model (2) is very useful. The proof is in the appendix.

Lemma 2.1 Suppose that (Y_t, λ_t) and (Y_t^m, λ_t^m) are defined by (2) and (3) respectively. If $0 \leqslant a + b < 1$, then the following statements hold:

1. $|\mathrm{E}\,(\lambda_t^m - \lambda_t)| = |\mathrm{E}\,(Y_t^m - Y_t)| \leqslant \delta_{1,m}$,
2. $\mathrm{E}\,(\lambda_t^m - \lambda_t)^2 \leqslant \delta_{2,m}$,
3. $\mathrm{E}\,(Y_t^m - Y_t)^2 \leqslant \delta_{3,m}$,

where $\delta_{i,m} \to 0$ as $m \to \infty$ for $i = 1, 2, 3$. Furthermore, almost surely, with m large enough

$$|\lambda_t^m - \lambda_t| \leqslant \delta \text{ and } |Y_t^m - Y_t| \leqslant \delta, \text{ for any } \delta > 0.$$

2.3 Non Linear Models

A simple generalization of the linear model (2) is given by

$$Y_t = N_t(\lambda_t), \;\; \lambda_t = \; f(\lambda_{t-1}) + b(Y_{t-1}), \tag{4}$$

for $t \geqslant 1$, and where $f(\cdot)$ and $b(\cdot)$ are known functions up to an unknown finite dimensional parameter vector. In addition both functions are defined and take values on the positive real line, that is $f, b : R^+ \to R^+$. The initial values Y_0 and λ_0 are fixed. It is seen that (2) is a special case of (4) upon defining $f(x) = d + ax$ and $b(x) = bx$, with $d, a, b > 0$, and $x \geqslant 0$.

There are many examples of non-linear time series models, see Tong (1990) and Fan and Yao (2003) for comprehensive reviews. Such models have not been considered in the literature earlier in the context of generalized linear models for count time series, and they provide a flexible framework for studying dependent count data. For example, consider the so called exponential autoregressive model which is described below.

Example 2.1 The exponential autoregressive model is defined by

$$Y_t = N_t(\lambda_t), \quad \lambda_t = \left(a + c\exp(-\gamma\lambda_{t-1}^2)\right)\lambda_{t-1} + bY_{t-1}. \tag{5}$$

The model parallels the structure of the traditional exponential autoregressive model, see Haggan and Ozaki (1981). Comparing recursions (4) and (5) it is clear that $f(x) = (a + c\exp(-\gamma x^2))\, x$ and $b(x) = bx$, $a, c, b, \gamma > 0$ and $x \geqslant 0$.

Reiterating previous arguments, notice that model (4) is related to a time series following generalized linear models as described in Kedem and Fokianos (2002). Specifically, for model (5), the Poisson assumption guarantees that the conditional random component of the model belongs to the exponential family of distributions, while the link function is equal to the identity, as before. If γ is known, then the systematic component of the model consists of the vector $(\lambda_{t-1}, \lambda_{t-1}\exp(-\gamma\lambda_{t-1}^2), Y_{t-1})'$; otherwise, (5) does not belong to the class of generalized linear models.

Assume the following conditions hold for $f(\cdot)$ and $b(\cdot)$:

Assumption NL

1. There exists a unique solution of the equation $\lambda = f(\lambda)$ and denote it by $\lambda^\star$.

2. With λ positive real and κ a positive integer, $f(\lambda)$ is increasing in λ for $\lambda > \lambda^\star$ and $b(\kappa)$ is increasing in κ such that $b(\kappa) \geqslant \beta^\star \kappa$, $\beta^\star > 0$.

3. For some $\alpha_2 > 0$, $f(\lambda_2) - f(\lambda_1) \leqslant \alpha_2(\lambda_2 - \lambda_1)$, for all $\lambda_1, \lambda_2 \geqslant \lambda^\star$ with $\lambda_2 \geqslant \lambda_1$.

4. For some $\beta_2 > 0$ such that $\alpha_2 + \beta_2 < 1$, $b(\kappa_2) - b(\kappa_1) \leqslant \beta_2(\kappa_2 - \kappa_1)$, $\kappa_2 \geqslant \kappa_1$.

To prove ϕ–irreducibility we again introduce an ϵ–perturbed model

$$Y_t^m = N_t\left(\lambda_t^m\right), \;\; \lambda_t^m = f(\lambda_{t-1}^m) + b(Y_{t-1}^m) + \varepsilon_{t,m} \;\; t \geqslant 1, \tag{6}$$

where $\varepsilon_{t,m}$ has been defined as in (3). The following proposition shows that the process $\{\lambda_t^m, t \geqslant 0\}$, as given by (6) is geometrically ergodic. Its proof parallels the proof of Proposition 2.1 and is given in the complete version of the manuscript at `web address`.

Proposition 2.3 Consider model (6) and suppose that Assumption NL holds true. Then the process $\{\lambda_t^m, t \geqslant 0\}$ is a geometrically ergodic Markov chain with finite moments of order k, for an arbitrary k.

In addition, we obtain the joint geometric ergodicity of the process $\{(Y_t^m, U_t, \lambda_t^m), t \geqslant 0\}$. The proof is omitted.

Proposition 2.4 Assume model (6) and suppose that the conditions of Proposition 2.3 hold true. Then the process $\{(Y_t^m, U_t, \lambda_t^m), t \geqslant 0\}$ is a $V_{(Y,U,\lambda)}$–geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y, U, \lambda) = 1 + Y^k + U^k + \lambda^k$.

We conclude this section with the following corollary for the perturbed exponential AR model

$$Y_t^m = N_t\left(\lambda_t^m\right), \;\; \lambda_t^m = \left(a + c\exp(-\gamma(\lambda_{t-1}^m)^2)\right)\lambda_{t-1}^m + bY_{t-1}^m + \varepsilon_{t,m} \;\; t \geqslant 1. \tag{7}$$

Corollary 2.1 Assume the exponential autoregressive model (7). Suppose that $0 < a + b < 1$. Then the process $\{(Y_t^m, U_t, \lambda_t^m), t \geqslant 0\}$ is a $V_{(Y,U,\lambda)}$–geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y, U, \lambda) = 1 + Y^k + U^k + \lambda^k$.

3 Likelihood Inference

Denote by θ the three dimensional vector of unknown parameters, that is $\theta = (d, a, b)'$ and the true value of the parameter as $\theta_0 = (d_0, a_0, b_0)'$. Then the conditional likelihood function for θ based on (2) and given the starting value λ_0 in terms of the observations $Y_1, \ldots, Y_n$ is given by

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

Here we have used the Poisson assumption, $\lambda_t(\theta) = d + a\lambda_{t-1}(\theta) + bY_{t-1}$ by (2) and $\lambda_t = \lambda_t(\theta_0)$. Hence, the log–likelihood function is given up to a constant, by

$$l_n(\theta) = \sum_{t=1}^{n} l_t(\theta) = \sum_{t=1}^{n} \left(y_t \log \lambda_t(\theta) - \lambda_t(\theta) \right), \tag{8}$$

and the score function is defined by

$$S_n(\theta) \;=\; \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n} \frac{\partial l_t(\theta)}{\partial \theta} = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\theta)} - 1 \right) \frac{\partial \lambda_t(\theta)}{\partial \theta}, \tag{9}$$

where $\partial\lambda_t(\theta)/\partial\theta$ is a three-dimensional vector with components given by

$$\frac{\partial \lambda_t}{\partial d} = 1 + a\frac{\partial \lambda_{t-1}}{\partial d}, \quad \frac{\partial \lambda_t}{\partial a} = \lambda_{t-1} + a\frac{\partial \lambda_{t-1}}{\partial a}, \quad \frac{\partial \lambda_t}{\partial b} = Y_{t-1} + a\frac{\partial \lambda_{t-1}}{\partial b}. \tag{10}$$

The solution of the equation $S_n(\theta) = 0$, if it exists, yields the conditional maximum likelihood estimator of θ which is denoted by $\hat{\theta}$. Furthermore, the Hessian matrix for model (2) is obtained by further differentiation of the score equations (9),

$$\begin{aligned} H_n(\theta) \;&=\; -\sum_{t=1}^{n} \frac{\partial^2 l_t(\theta)}{\partial\theta\partial\theta'} \\ &=\; \sum_{t=1}^{n} \frac{Y_t}{\lambda_t^2(\theta)} \left(\frac{\partial \lambda_t(\theta)}{\partial \theta} \right) \left(\frac{\partial \lambda_t(\theta)}{\partial \theta} \right)' - \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\theta)} - 1 \right) \frac{\partial^2 \lambda_t(\theta)}{\partial\theta\partial\theta'}. \end{aligned} \tag{11}$$

If the process $\{(Y_t, \lambda_t)\}$ is a geometrically ergodic Markov chain, then an asymptotic theory for the maximum likelihood estimator of θ can be developed directly. A problem is that at present

we do not know precisely what conditions guarantee ergodicity of (2). However, the assumptions of Proposition 2.2 guarantee geometric ergodicity of the perturbed model (Y_t^m, λ_t^m). In addition, Lemma 2.1 shows that λ_t^m approaches λ_t, for large m. Hence, it is rather natural to use the ergodic properties of the perturbed process (Y_t^m, λ_t^m) to study the asymptotic properties of the maximum likelihood estimators analogous to (9) and then use Lemma 2.1. Towards this goal, we define the counterparts of expressions (8)-(11) for model (3).

The likelihood function, say L^m, including the pseudo observations $U_1, U_2, \ldots, U_n$, is given by

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

by the Poisson assumption and the asserted independence of U_t from $(Y_{t-1}^m, \lambda_{t-1}^m)$. Here, $f_u(\cdot)$ denotes the uniform density and $\lambda_t^m(\theta) = d + a\lambda_{t-1}^m(\theta) + bY_{t-1}^m + \varepsilon_{t,m}$ as given by (3). We note that $L(\theta)$ and $L^m(\theta)$ have identical form with the only exception that (Y_t, λ_t) is replaced by (Y_t^m, λ_t^m). Therefore, $S_n^m(\theta)$ and $H_n^m(\theta)$ have the same form as (9) and (11)–with recusrsions defined by (10)–but with (Y_t, λ_t) replaced by (Y_t^m, λ_t^m). The solution of the equation $S_n^m(\theta) = 0$ is denoted by $\hat{\theta}^m$. See `web adress`, for a more detailed exposition.

3.1 Asymptotic Theory

To study the asymptotic properties of the maximum likelihood estimator $\hat{\theta}$, for the linear model (2) we derive and use the asymptotic properties of the maximum likelihood estimator $\hat{\theta}^m$ for the perturbed linear model (3). The main tool in linking $\hat{\theta}$ to $\hat{\theta}^m$ is Prop. 6.3.9 of Brockwell and Davis (1991). Accordingly, we first show that $\hat{\theta}^m$ is asymptotically normal where for the proof of consistency and asymptotic normality we take advantage of the fact that the log–likelihood function is three times differentiable applying Jensen and Rahbek (2004, Lemma 1). Then we show that the score function, the information matrix and the third derivatives of the perturbed likelihood function tend to the corresponding quantities of the unperturbed likelihood function (8) which enable us to employ Brockwell and Davis (1991, Prop. 6.3.9). To formulate the end result, introduce lower and upper values of each component of θ, $\delta_L < d_0 < \delta_U$, $\alpha_L < a_0 < \alpha_U < 1$ and $\beta_L < b_0 < \beta_U$, and

in terms of these define,

$$O\left(\theta_0\right) = \left\{\theta | 0 < \delta_L \leqslant d \leqslant \delta_U,\ \ 0 < \alpha_L \leqslant a \leqslant \alpha_U < 1 \ \text{ and } \ 0 < \beta_L \leqslant b \leqslant \beta_U\right\}. \tag{12}$$

Then the following theorem regarding the properties of the maximum likelihood estimator $\hat{\theta}$ holds true.

Theorem 3.1 Consider model (2) and suppose that at the true value θ_0, $0 < a_0 + b_0 < 1$. Then, there exists a fixed open neighborhood $O = O(\theta_0)$ of θ_0–see (12)–such that with probability tending to one, as $n \to \infty$, the log likelihood function (8) has a unique maximum point $\hat{\theta}$. Furthermore, $\hat{\theta}$ is consistent and asymptotically normal,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}(0, G^{-1}),$$

where the matrix G is defined in Lemma 3.1. A consistent estimator of G is given by $G_n(\hat{\theta})$, where

$$G_n(\theta) = \sum_{t=1}^{n} \text{Var}\left[\frac{\partial l_t(\theta)}{\partial \theta} \mid \mathcal{F}_{t-1}\right] \ = \ \sum_{t=1}^{n} \frac{1}{\lambda_t(\theta)} \left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right) \left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right)'$$

To prove the above theorem, we need a series of results. Lemma 3.1 shows that the limiting conditional information matrix of model (3) tends to another matrix that plays the role of the conditional information matrix for model (2). In Lemmas 3.2–3.4 the conditions (A.1), (A.2) and (A.3) of Lemma 1 of Jensen and Rahbek (2004) are verified for the perturbed model (3). In addition, these lemmas show that the score function, the Hessian matrix and the third derivative of the log likelihood function of the perturbed model tend to their counterparts of the unperturbed model. The proof of Lemma 3.2 which illustrates the applied technique, is in the Appendix. The other proofs can be found in the complete version of the manuscript at `web address`.

Lemma 3.1 Define the matrices

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

Under the assumptions of Theorem 3.1, the above matrices evaluated at the true value $\theta = \theta_0$ satisfy $G^m \to G$, as $m \to \infty$. In addition, G^m and G are positive definite.

Lemma 3.2 Under the assumptions of Theorem 3.1, the score functions defined by (9) and its counterpart for the perturbed model (3) and evaluated at the true value $\theta = \theta_0$ satisfy the following:

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

Lemma 3.3 Under the assumptions of Theorem 3.1, the Hessian matrices defined by (11) and and its counterpart for the perturbed model (3) and evaluated at the the true value $\theta = \theta_0$ satisfy the following:

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

Lemma 3.4 With the neighborhood $O(\theta_0)$ defined in (12), it holds under the assumptions of Theorem 3.1, that,

$$\max_{i,j,k=1,2,3} \sup_{\theta\in O(\theta_0)} \left| \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^3 l_t(\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \right| \leqslant M_n := \frac{1}{n} \sum_{t=1}^{n} m_t$$

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

$$\begin{aligned} m_t &= C\left(Y_t\mu_{3t} + \mu_{3t} + Y_t\mu_{2t}\mu_{1t} + Y_t\mu_{1t}^3\right) \\ \mu_{it} &= \beta_U \sum_{j=1}^{t-i} k_{j,i}\alpha_U^{j-1} Y_{t-i-j}, \quad k_{j,1} = j,\ k_{j,2} = j(j+1) \text{ and } k_{j,3} = j(j+1)(j+2). \end{aligned}$$

Define correspondingly M_n^m, m_t^m and μ_{it}^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. $\lim_{m\to\infty} \limsup_{n\to\infty} P\left(|M_n^m - M_n| > \varepsilon n\right) = 0$, for every $\varepsilon > 0$.

3.2 Some Remarks about the Exponential Autoregressive Model

Study of asymptotic properties of the maximum likelihood estimator for the non linear model (6) proceeds along the previous lines but the corresponding analysis is more cumbersome. For instance, consider the exponential autoregressive model (5) and put $\theta = (a, c, b)'$ and assume that the parameter γ is known. Then the recursions for calculation of the score are given by

$$\begin{aligned}
\frac{\partial \lambda_t}{\partial a} &= \left(1 - 2\gamma c \lambda_{t-1} \exp(-\gamma \lambda_{t-1}^2) \frac{\partial \lambda_{t-1}}{\partial a}\right) \lambda_{t-1} + \left(a + c \exp(-\gamma \lambda_{t-1}^2)\right) \frac{\partial \lambda_{t-1}}{\partial a}, \\
\frac{\partial \lambda_t}{\partial c} &= \left(1 - 2\gamma c \lambda_{t-1} \frac{\partial \lambda_{t-1}}{\partial c}\right) \exp(-\gamma \lambda_{t-1}^2) \lambda_{t-1} + \left(a + c \exp(-\gamma \lambda_{t-1}^2)\right) \frac{\partial \lambda_{t-1}}{\partial c}, \\
\frac{\partial \lambda_t}{\partial b} &= a \frac{\partial \lambda_{t-1}}{\partial b} + \left(1 - 2\gamma \lambda_{t-1}^2\right) c \exp(-\gamma \lambda_{t-1}^2) \frac{\partial \lambda_{t-1}}{\partial b} + Y_{t-1}.
\end{aligned}$$

If the parameter γ is assumed to be unknown, then the following additional recursion is needed to obtain the score equations for the enlarged vector of parameters which includes γ:

$$\frac{\partial \lambda_t}{\partial \gamma} = -c \exp(-\gamma \lambda_{t-1}^2) \lambda_{t-1}^2 \left(\lambda_{t-1} + 2\gamma \frac{\partial \lambda_{t-1}}{\partial \gamma}\right) + \left(a + c \exp(-\gamma \lambda_{t-1}^2)\right) \frac{\partial \lambda_{t-1}}{\partial \gamma}.$$

We remark that when $c = 0$ the parameter γ is not identifiable. Therefore testing the hypothesis that $c = 0$ by means of the likelihood ratio test, for instance, is more involved but it can be accomplished by means of the methodology outlined by Davies (1987). Although we do not study in detail the asymptotic properties of the maximum likelihood estimator for the exponential autoregressive model, we report some simulation evidence in the next section–for more see `web address`.

4 Simulation and Data Analysis

A few simulation results are reported for illustration of the asymptotic normality of the maximum likelihood estimator for both the linear and non linear model. Calculation of the maximum likelihood estimators is carried out by optimizing the log–likelihood function (8) employing a quasi–Newton method–details are available by the authors. The simple linear model (2) is generated with $(d_0, a_0, b_0) = (0.3, 0.4, 0.5)$ for different sample sizes. For this choice of parameter values, $a_0 + b_0 = 0.9$ which is close to the sum $\hat{a} + \hat{b}$ of the estimates obtained from the real data analysis reported in Section 4.3. Results for different set of parameter values were similar, provided that the condition $a_0 + b_0 < 1$ is satisfied, and therefore we omit them.

4.1 Simulations for the Linear Model

In Ferland et al. (2006) it was shown that (2) possesses moments up to second order identical to those of an ARMA(1,1) process which satisfies the difference equation

$$(Y_t - \mu) - (a + b)(Y_{t-1} - \mu) = e_t - ae_{t-1},$$

where e_t is a white noise process with mean and variance equal to $d/(1 - a - b)$. This observation points to the potential use of the least squares theory for estimation of the parameter vector (d, a, b). We compare the least squares estimators with the maximum likelihood estimators in what follows. To initiate the algorithm for optimization of the log–likelihood function (8) using the recursions (10), starting values for (d, a, b) are obtained by the ARIMA(1,1) fit to the data.

Table 1 shows the results of the comparison between maximum likelihood and least squares estimators. The table reports the estimates of the parameters obtained by averaging out the results from all runs together with their mean square error (in parentheses). The mean square error has been calculated by using the simulation output. Notice that in all cases considered the maximum estimators have lower mean square error than the least squares estimators and as the sample sizes increases the error of both type of estimators decreases. In addition Table 2 compares the standard errors of the estimators obtained by the standard deviation of the simulated estimators and by the

asymptotic theory, as developed in Theorem 3.1 and Lemma 3.1. The results show that the standard errors based on the simulations are somewhat larger than the ones obtained after calculating and inverting the matrix G but this difference gradually become smaller as the sample size increases. Figure 1 shows histograms and qq-plots for the sampling distribution of the standardized maximum likelihood estimators indicating the adequacy of the normal approximation although for a sample size of 500 there are some deviations in the upper plot for $\hat{d}$. It is seen from Figure 1 that $\hat{d}$ is plagued by some extreme values in the right tail. Its inferior accuracy compared to $\hat{a}$ and $\hat{b}$ is also apparent from Table 2. This behavior is easier to explain from the least squares estimates. From the defining equation (2), it is seen that the least square estimator of d will be sensitive to the extreme right tail events of the distribution of Y_t and this is reflected in Figure 1. The estimates of a and b on the other hand, are ratio type estimators where extreme events are balanced out in the numerator and denominator. For larger sample sizes, the approximation to normality is improving.

Sample Size	Maximum Likelihood Estimators			Least Squares Estimators		
n	d	a	b	d	a	b
200	0.3713	0.3756	0.4967	0.3909	0.3790	0.4863
	(0.0254)	(0.0094)	(0.0056)	(0.0334)	(0.0252)	(0.0257)
500	0.3271	0.3923	0.4971	0.3318	0.3922	0.4932
	(0.0071)	(0.0030)	(0.0196)	(0.0100)	(0.0136)	(0.0159)
1000	0.3148	0.3954	0.4985	0.3180	0.3951	0.4965
	(0.0027)	(0.0014)	(0.0009)	(0.0043)	(0.0107)	(0.0130)

Table 1: Estimators and their mean square error (in parentheses) for model (2) when $(d_0, a_0, b_0) = (0.3, 0.4, 0.5)$ and for different sample sizes by both maximum likelihood and least squares methods. Results are based on 1000 simulations.

Sample Size	Simulated standard errors			Standard Errors from $G(\theta_0)$		
n	d	a	b	d	a	b
200	0.1429	0.0940	0.0749	0.0937	0.0733	0.0593
500	0.0803	0.0548	0.0443	0.0574	0.0459	0.0372
1000	0.0505	0.0380	0.0314	0.0403	0.0323	0.0263

Table 2: Comparison of standard errors for model (2) with $(d_0, a_0, b_0) = (0.3, 0.4, 0.5)$.

4.2 Simulations for the Exponential Autoregressive Model

Table 3 shows some empirical results from estimation of the exponential autoregressive model (5) with $(a_0, c_0, b_0) = (0.25, 1, 0.65)$ and for different values of γ. Note that the standard errors have been obtained by the standard deviation of the simulated estimators, as in the case of the linear model. Initially it is assumed that the parameter γ is known. Maximization of the log likelihood function (8) under (5) is carried out as it was discussed in the case of the linear model. Initial values for the parameters a and b are obtained by the fit of a linear model (2), while the initial value of c is fixed at 0.50. The results of Table 3 show that for large sample sizes, the maximum likelihood estimators are approaching the true parameter values when the parameter γ is known.

Application of the exponential autoregressive model (5) becomes more involved when the parameter γ is unknown, which is the most interesting case in applications. To estimate jointly the parameter vector (a, c, b, γ) of (5) the following procedure is proposed.

- Fit the linear model (2) to the data to obtain starting values for both a and b.
- Set the initial value of c equal to some constant.
- Generate a grid of values for γ and for each of these values fit model (5) with known γ.
- To maximize the log likelihood function over all (a, c, b, γ), get as a starting value the γ–value that yields the maximum log–likelihood from the previous step together with the corresponding coefficients.

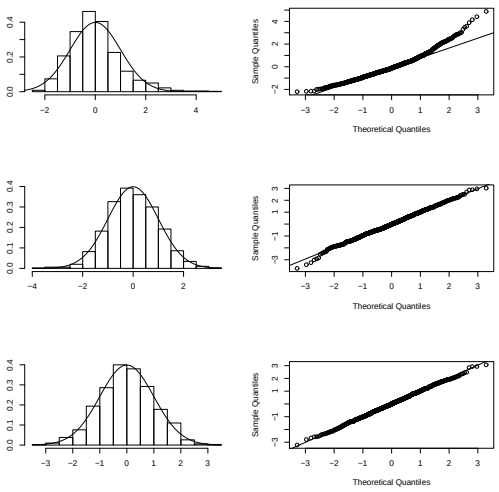

Figure 1: From top to bottom: Histograms and qq-plots of the sampling distribution of the standardized estimators of $\hat{\theta} = (\hat{d}, \hat{a}, \hat{b})$ for the linear model (2) when the true values are $(d_0, a_0, b_0) = (0.3, 0.4, 0.5)$. Superimposed is the standard normal density function. The results are based on 500 data points and 1000 simulations.

Maximization is carried out numerically as described before. Table 4 reports the results of the above method and it is observed that the sample size should be reasonably large for an adequate approximation of both c and γ.

Maximum Likelihood Estimators			Sample Size	
$\hat{a}$	$\hat{c}$	$\hat{b}$	n	γ
0.2470	1.0241	0.6468	500	2
(0.0465)	(0.1282)	(0.0472)		
0.2486	1.0128	0.6475	1000	
(0.0339)	(0.0898)	(0.0344)		
0.2504	1.0300	0.6440	500	1
(0.0472)	(0.1416)	(0.0473)		
0.2488	1.0140	0.6484	1000	
(0.0336)	(0.0943)	(0.0324)		
0.2510	1.0262	0.6457	500	0.50
(0.0496)	(0.1429)	(0.0469)		
0.2517	1.0106	0.6465	1000	
(0.0327)	(0.0987)	(0.0313)		

Table 3: Estimates and their standard error (in parentheses) for the exponential autoregressive model (5) when $(a_0, c_0, b_0) = (0.25, 1, 0.65)$ and for different sample sizes by maximum likelihood. The parameter γ is assumed to be known. Results are based on 1000 simulations.

Maximum Likelihood Estimators				Sample Size	True γ
$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{\gamma}$	n	γ
0.2504	1.1645	0.6430	1.8655	500	1.50
(0.0537)	(0.5348)	(0.0487)	(1.2573)		
0.2491	1.0593	0.6456	1.6357	1000	
(0.0364)	(0.2498)	(0.0343)	(0.7684)		
0.2500	1.1190	0.6426	1.1542	500	1
(0.0521)	(0.4431)	(0.0485)	(0.6919)		
0.2490	1.0278	0.6475	1.0325	1000	
(0.0355)	(0.1690)	(0.0334)	(0.2886)		
0.2467	1.1025	0.6472	0.5419	500	0.50
(0.0514)	(0.3813)	(0.0463)	(0.2341)		
0.2481	1.0322	0.6493	0.5064	1000	
(0.0337)	(0.1590)	(0.0305)	(0.1087)		

Table 4: Estimates and their standard errors (in parentheses) for the exponential autoregressive model (5) when $(a_0, c_0, b_0, \gamma_0) = (0.25, 1, 0.65, \gamma_0)$ where $\gamma_0 \in \{0.5, 1, 1.5\}$ and for different sample sizes by maximum likelihood. Results are based on 500 simulations.

4.3 Data Example

For an illustration of the methodology, models (2) and (5) are applied to real data which consist of the number of transactions per minute for the stock Ericsson B during July 2nd 2002. This is a part of a larger data set which includes all the transactions of this specific stock for the time period between July 2nd and July 22nd, 2002. There are 460 available observations conveying eight hours of transactions, approximately. Notice that the first and last minutes transactions are not taken into account. Figure 2 shows both the data and the respective autocorrelation function. Even though the data are counts, the plot of the usual autocorrelation functions reveals the high dependence between transactions. Note that the mean number of transactions for these particular data is equal to 9.909 while their sample variance is given by 32.836. This is a case of overdispersion, as it was discussed in Section 2.1. To model these data, set $\lambda_0 = 0$ and $\partial\lambda_0/\partial\theta = 0$ for initialization of

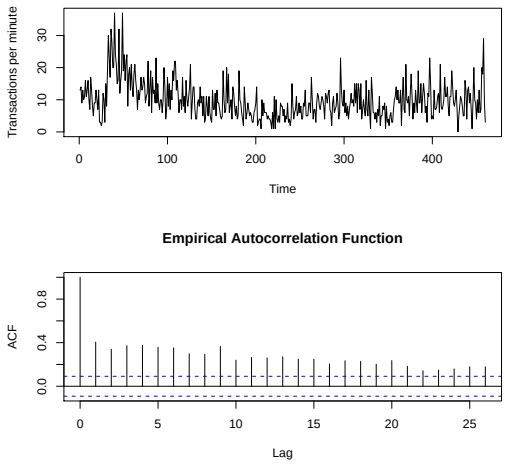

Figure 2: Number of transactions per minute for the stock Ericsson B during July 2nd, 2002. The bottom plot shows their autocorrelation function

the recursions and consider the linear model (2). Maximization of the log–likelihood function (8)

yields the following results:

$$\begin{array}{lcccccc} \hat{\lambda}_t = & 0.5808 & +0.7445 & \hat{\lambda}_{t-1} + & 0.1986 & Y_{t-1} \\ & (0.1628) & (0.0264) & & (0.0167) & \end{array}$$

where the standard errors underneath the estimated parameter are computed by using the so called robust sandwich matrix $H_n(\hat{\theta})G_n^{-1}(\hat{\theta})H_n(\hat{\theta})$, where $G_n(\theta)$ has been defined in Theorem 3.1 and $H_n(\theta)$ is given by (11). More precisely, the latter matrix is inverted and the standard errors are computed by taking a square root along its diagonal.

It is rather interesting to observe that $\hat{a} + \hat{b}$ is close to unity. In fact, an approximate 95% confidence interval for the parameter $a + b$ is given by $(0.9097, 0.9765)$. This is a very similar phenomenon to the unit-root case in autoregressive time series in econometric analysis of most financial data, and moreover, similar to the IGARCH literature where high-persistence is often discovered in the conditional variance. Similar findings can be derived from Rydberg and Shephard (2000) where transaction data were also examined.

To examine the adequacy of the fit, consider the so called Pearson residuals defined by $e_t = (Y_t - \lambda_t)/\sqrt{\lambda_t}$. Under the correct model, the sequence e_t is a white noise sequence with constant variance, see Kedem and Fokianos (2002, Sec. 1.6.3). To estimate the Pearson residuals, substitute λ_t by $\lambda_t(\hat{\theta})$. Figure 3 demonstrates that the predicted values defined by $\hat{Y}_t = \lambda_t(\hat{\theta})$ approximate the observed process reasonably well. The bottom plot of the same figure illustrates the whiteness of the Pearson residuals by depicting the cumulative periodogram plot, see Brockwell and Davis (1991, Sec. 10.2).

Consider now the application of the Poisson exponential autoregressive model (5) to the same data. The following model is fitted

$$\begin{array}{lccccccc} \hat{\lambda}_t = & (0.8303 & +7.030 & \exp(-0.1675\hat{\lambda}_{t-1}^2)) & \hat{\lambda}_{t-1} & + & 0.1551 & Y_{t-1} \\ & (0.0232) & (3.0732) & (0.0592) & & & (0.0218) & \end{array}$$

where the stated standard errors refer to those of $\hat{a}$, $\hat{c}$ $\hat{d}$ and $\hat{\gamma}$, respectively. The starting value for carrying out the estimation of γ is implemented by the suggested profiling procedure. To be more specific, a grid of values is generated and the corresponding log–likelihood value is evaluated at

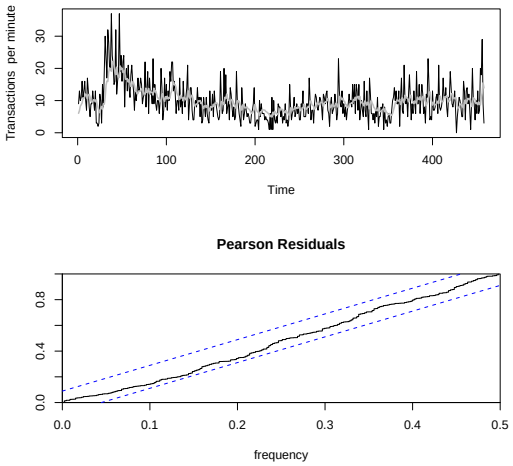

Figure 3: Top: Observed and predicted (grey) number of transactions per minute using (2). Bottom: Cumulative periodogram plot of the Pearson residuals.

these points. For these data, the values of the parameter γ were generated from 0.001 to 10 by increments of 0.20, that is 50 values were obtained. The upper plot of Figure 4 shows a graph of γ versus the corresponding log–likelihood function for a part of the generated grid, and it shows that at the value 0.20506, approximately, the corresponding maximum value is obtained. Using this value as a starting point for γ, model (5) is fitted to the data. The other graphs of Figure 4 show the predicted response–notice again that $\hat{Y}_t = \hat{\lambda}_t$–and the cumulative periodogram plot of the Pearson residuals. To compare the models, we calculate the mean square error of the Pearson residuals defined by $\sum_{t=1}^{N} e_t^2/(N-p)$, where p is the number of estimated parameters, see Kedem and Fokianos (2002, Sec. 1.8). It turns out that for the linear model the mean square error of the Pearson residuals is equal to 2.3686 while for the non linear model it is equal to 2.3923–that means both of the models yield similar conclusions–see also Figure 5 which depicts the Pearson residuals from both models applied to these data. As a final remark, we note that the numerical results, in particular for $\hat{d}$, are sensitive to the choice of the initial value λ_0. Similar problems are encountered in the GARCH model fitting where it is well known that different software might produce different

results.

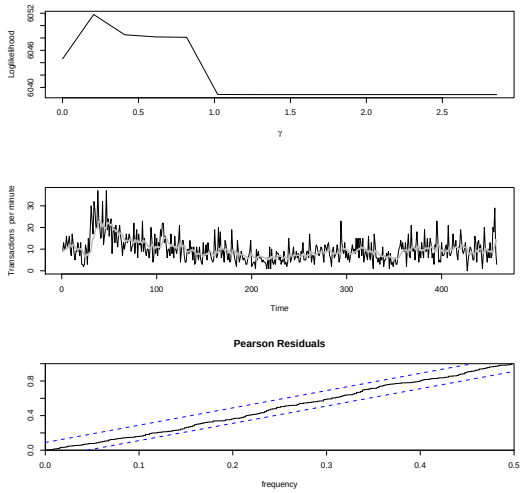

Figure 4: Top: Obtaining an estimator for γ. Center: Observed and predicted (grey) number of transactions per minute using (5). Bottom: Cumulative periodogram plot of the raw residuals.

Acknowledgements

The authors acknowledge constructive criticism and helpful remarks from the Associate Editor and two reviewers. Many thanks also to the participants at the CFE08 conference and seminar participants at LSE and Bologna University. Comments from Neil Shephard, Andrew Patton and Theis Lange are greatly acknowledged. K. Brännäs provided us with the transactions data. A. Rahbek would like to thank for financial support from Danish Social Sciences Research Council, Project Number 2114-04-001.

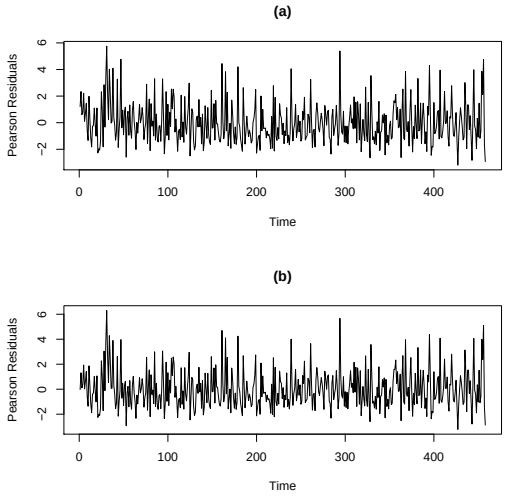

Figure 5: (a) Pearson residuals when model (2) is applied to transactions data. (b) Pearson residuals when model (5) is applied to transactions data.

Appendix

Recall that $\lambda^* = d/(1-a)$ is a fix point of the skeleton $\lambda_t = d + a\lambda_{t-1}$ of (2) and (3). We start by proving that $\{\lambda_t\}$, defined by (2) or (3), is open set irreducible on $[\lambda^*, \infty)$. The proof of the following lemma does not require the ϵ-perturbation of (3), but it is valid in the presence of such a perturbation.

Lemma A-1 Let $\{\lambda_t\}$ be a Markov chain defined by (2) or (3). If $0 < a < 1$, then every point in $[\lambda^*, \infty)$ is reachable.

Proof: To simplify notation, we state the proof for model (2). Consider a point $c \in [\lambda^*, \infty)$. We may assume that $\lambda_1 > c$, since otherwise we may start from $\lambda_j = d + a\lambda_{j-1} + bY_{j-1} > \lambda^*$. Consider a path such that $Y_1 = \ldots = Y_i = 0$. Then λ_i approaches λ^* as i increases and $a < 1$ and this proves that c is reachable if $c = \lambda^*$. (This was already proved by Streett (2000)). The proof is more difficult if $c > \lambda^*$. The intuitive idea is to consider realizations where $Y_1 = N$, and where subsequent $\{Y_i, i > 1\}$ are zero. Then, if (c_1, c_2) is an open interval containing c, λ_i will

approach this interval from above in successively smaller steps. By choosing N appropriately, one of these steps will be contained in (c_1, c_2). Equivalently, we can prove that one of these steps will be arbitrarily close to c. Therefore, let $Y_1 = N$ and $Y_2 = 0$, $Y_3 = 0, \ldots$. Then, for $j \geqslant 1$

$$\lambda_{1+j} = \lambda_{1+j}(N) = a^{j-1}(a\lambda_1 + bN) + \frac{1-a^j}{1-a}d.$$

Let $N := N_j$ be the least integer such that

$$\lambda_{1+j}(N) > c, \quad \lambda_{1+j}(N-1) \leqslant c.$$

(Note that we can always choose λ_1 such that $N - 1 > 1$ so that the reasoning is exactly the same for the ϵ–perturbed chain given in (3)). We have $\lambda_{1+j}(N) - \lambda_{1+j}(N-1) = a^{j-1}b$. For any $\delta > 0$, since $a < 1$, we can choose a j such that $a^{j-1}b < \delta$, that is $c \in [\lambda_{1+j}(N-1), \lambda_{1+j}(N)]$ where the width of the interval is less than δ. Since δ is arbitrary, c can be approximated arbitrarily closely, and c is reachable if $c \in [\lambda^*, \infty)$. □

To prove Proposition 2.1 we need to show that $\{\lambda_t^m, t \geqslant 0\}$ is aperiodic and ϕ–irreducible. In addition we show the existence of a small set C and a test function $V(\cdot)$ which satisfies

$$\mathrm{E}[V(\lambda_{t+1}^m)|\lambda_t^m = \lambda] \leqslant (1-k_1)V(\lambda) + k_2 1(\lambda \in C) \tag{A-1}$$

for some constants k_1, k_2 such that $0 < k_1 < 1$, $0 < k_2 < \infty$. This implies that the chain $\{\lambda_t^m, t \geqslant 0\}$ is geometrically ergodic and as will be seen, with a proper choice of V, that the k'th moment of λ_t^m exists for an arbitrary k, see Meyn and Tweedie (1993). The inequality (A-1) and uniform open set reachability from a compact set $[\lambda^*, K]$, $K > \lambda^*$, can be established for both (2) and (3), but the ϵ–perturbation is used to establish ϕ–irreducibility and uniform ϕ–reachability required to establish the existence of a small set.

Proof of Proposition 2.1

From Lemma A-1 we have that $\{\lambda_t^m, t \geqslant 0\}$ as defined by (3) is open set irreducible on $[\lambda^*, \infty)$. Let A be a set in the support of $[k, \infty)$ of ϕ for some $k > \lambda^*$, where ϕ is the Lebesgue measure,

and such that $\phi(A) > 0$. Let $c^{'}$ be a point in A. Then using the technique of proof of Lemma A-1, for some j, λ_{j+1}^m will be arbitrarily close to $(c^{'} - d - b)/a$, where $(c^{'} - d - b)/a > \lambda^*$ by choosing k large enough. In particular, j can be chosen so that $\mid d + a\lambda_{j+1}^m + b - c^{'} \mid < \epsilon/2$. Therefore if $D = A \cap B$ with $B = (c^{'} - \delta/2, c^{'} + \delta/2)$ for some small δ, and $f_u(\cdot)$ the density of U_t, then the probability of being in A in the next step is

$$P(A) \geqslant P(D) = \int_D f_u(u)du \geqslant \inf_D f_u(u)\phi(D) > 0,$$

which implies ϕ-irreducibility. (Note that it follows from this proof that the ϵ–perturbation need not be introduced for $Y = 1$ but may in fact be inserted for any Y.) It remains to prove the existence of a small set, aperiodicity and the inequality (A-1).

Existence of a small set can be proved by extending and modifying the technique of Lemma A-1. Let C be a compact set, $C = [\lambda^*, K]$ for a finite $K > \lambda^*$. Since $a < 1$ and K is finite, there exists an integer $n = n(\eta)$ such that for given $\eta > 0$, and with a path where $Y_1 = \cdots = Y_{n-1} = 0$, $\lambda_1^m = \lambda$, $|\lambda_n^m - \lambda^*| = a^n|\lambda - \lambda^*| < \eta$ for all $\lambda \in C$. Then with $Y_n^m = N$, $Y_{n+1}^m = 0, Y_{n+2}^m = 0, \ldots$,

$$\lambda_{n+j}^m = a^j\lambda_n^m + a^{j-1}bN + \frac{1-a^j}{1-a}d = a^j(\lambda_n^m - \lambda^*) + a^{j-1}bN + \lambda^*.$$

Similarly to the proof of Lemma A-1, consider an open interval (c_1, c_2) with $c_1 > \lambda^*$, and let N be the least integer such that

$$\mu_{n+j}(N) \equiv a^{j-1}bN + \lambda^* > c_2,$$

where with no loss of generality we may assume that $N > 2$. By choosing j large enough and by the previous arguments, for any $\delta > 0$, there exists an j such that $a^{j-1}b < \delta$. Therefore

$$c_2 - \delta < \mu_{n+j}(N-1) < c_2.$$

But $\lambda_{n+j}^m(N-1) = \mu_{n+j}(N-1) + a^j(\lambda_n^m - \lambda^*)$ where by choosing η small enough

$$c_2 - \delta < \mu_{n+j}(N-1) < \lambda_{n+j}^m(N-1) < c_2$$

so that for these choices of n, j, N, $\lambda_{n+j}^m(N-1) \in (c_1, c_2)$ for all $\lambda \in C$ and

$$\inf_{\lambda \in C} \mathrm{P}^{n+m}(\lambda, (c_1, c_2)) \geqslant \mathrm{P}(Y_1^m = 0, \ldots Y_{n-1}^m = 0, Y_n^m = N, Y_{n+1}^m = 0, \ldots, Y_{n+j-1}^m = 0) > 0.$$

This means that the interval (c_1, c_2) is uniformly reachable from all $\lambda \in [\lambda^*, K]$, and arguing as in the above proof of ϕ–irreducibility it follows that an n can be found such that

$$\inf_{\lambda \in C} \mathrm{P}^n(\lambda, A) > 0,$$

for a set A of positive Lebesgue measure. This implies that the set C is a small set.

We now show that $\{\lambda_t^m, t \geqslant 0\}$ is aperiodic. Consider the small set $C = [\lambda^*, K]$. Note that $\phi(C) > 0$ and let $\lambda_{t-1}^m = \lambda \in C$. Then $\lambda_t^m = d + a\lambda + b\varepsilon_{t,m}$. If $Y_{t-1}^m = 0$, then $\lambda_t^m = d + a\lambda = \lambda^*(1-a) + a\lambda = \lambda^* + a(\lambda - \lambda^*) \geqslant \lambda^*$ since $a > 0$. On the other hand $\lambda_t^m - \lambda = \lambda^*(1-a) - \lambda(1-a) = -(1-a)(\lambda - \lambda^*) \leqslant 0$ for $a < 1$. It is concluded that if $0 < a < 1$, then $\lambda \in C$ and $Y_{t-1}^m = 0$ imply that $\lambda_t \in C$, which in turn implies $\mathrm{P}(\lambda, C) \geqslant \mathrm{P}(Y_{t-1}^m = 0 \mid \lambda_{t-1} = \lambda) = \mathrm{P}(N_t(\lambda) = 0) > 0$. Similarly, $\mathrm{P}^2(\lambda, C) \geqslant \mathrm{P}(Y_t^m = Y_{t-1}^m = 0 \mid \lambda_{t-1} = \lambda) > 0$. It follows that $\{\lambda_t^m, t \geqslant 0\}$ is aperiodic by Chan (1990, Prop. A1.1).

Finally, we prove the existence of a test function $V(\cdot)$ such that (A-1) holds true. Consider $V(x) = 1 + x^k$. Then

$$\begin{aligned}
\mathrm{E}[V(\lambda_t^m)|\lambda_{t-1}^m = \lambda] &= \mathrm{E}[(1 + (\lambda_t^m)^k)|\lambda_{t-1} = \lambda] \\
&= 1 + \mathrm{E}[(d + a\lambda + bY_{t-1}^m + \varepsilon_{t,m})^k|\lambda_{t-1} = \lambda] \\
&= 1 + \sum_{i=0}^{k} \binom{k}{i} (a\lambda)^i (b\lambda)^{k-i} + \sum_{j=0}^{k-1} c_j \lambda^j \\
&= 1 + (a+b)^k \lambda^k + \sum_{j=0}^{k-1} c_j \lambda^j,
\end{aligned}$$

for some constants c_j depending on a, b, d and ϵ. Consider the small set $C = [\lambda^*, K]$ and write

$$1 + (a+b)^k \lambda^k = \left[1 - \frac{\lambda^k[1 - (a+b)^k]}{1 + \lambda^k}\right] (1 + \lambda^k) \left[1(\lambda \in C) + 1(\lambda \in C^c)\right].$$

For $\lambda \in C^c$, we obtain

$$\sup_{\lambda \in C^c} \left[1 - \frac{\lambda^k[1 - (a+b)^k]}{1 + \lambda^k}\right] \to 1 - [1 - (a+b)^k] = (a+b)^k$$

as K increases. Similarly, by making K large enough

$$\sup_{\lambda \in C^c} \frac{\sum_{j=1}^{k-1} c_j \lambda^j}{1+\lambda^k} < \delta,$$

where $0 < \delta < (a+b)^k$. But on C, $1 + (a+b)^k \lambda^k + \sum_{j=1}^{k-1} c_j \lambda^j$ is bounded. Therefore, it follows that there exists constants k_1 and k_2 such that $(0 < k_1 < 1, 0 < k_2 < \infty)$

$$\mathrm{E}(V(\lambda_t^m) | \lambda_{t-1}^m = \lambda) \leqslant (1-k_1)V(\lambda) + k_2 1(\lambda \in C)$$

and this implies that the chain $\{\lambda_t^m, t \geqslant 0\}$ is geometrically ergodic, and that the kth moment of λ_t^m exists for an arbitrary k.

Proof of Proposition 2.2

We will use the method of Meitz and Saikonnen (2008) to show that geometric ergodicity of the $\{\lambda_t^m\}$ process implies geometric ergodicity of the chain $\{(Y_t^m, U_t, \lambda_t^m)\}$, defined by (3). Denote the σ-algebra generated by the past of U_{t+1} and $N_t(\cdot)$ process by $\mathcal{F}_{t-1}$, that is $\mathcal{F}_t = \sigma(U_{k+1}, N_k, k \leqslant t)$.

First note that conditional distribution of Y_t^m given $\mathcal{F}_{t-1}$ depends only on λ_t^m. In addition, the conditional distribution function of Y_t^m given $\lambda_t^m = \lambda$ does not depend on t. Further, given the initial state $(Y_0^m, \lambda_0^m, U_1)$, we have

$$\lambda_1^m = d + a\lambda_0^m + bY_0^m + \varepsilon_{1,m}$$

and since the conditional distribution of Y_1^m given $\{Y_0^m, U_1, \lambda_0^m, \lambda_1^m\}$ is Poisson(λ_1), conditionally on $\{Y_0^m, U_1, \lambda_0^m, \lambda_1^m\}$ we have

$$\lambda_2^m = d + a\lambda_1^m + bN_1(\lambda_1^m) + \varepsilon_{2,m}$$

where $N_1(\lambda_1)$ is Poisson(λ_1). Hence, $\{\lambda_t^m\}$ considered as component of the trivariate chain $\{(Y_t^m, U_t, \lambda_t^m\}$ from $t \geqslant 2$ has the same structure as the one–dimensional chain $\{\lambda_t^m\}$, where conditionally on $\mathcal{F}_{t-1}$

$$\lambda_t^m = d + a\lambda_{t-1}^m + bN_{t-1}(\lambda_{t-1}^m) + \varepsilon_{t,m}$$

with $N_{t-1}(\lambda_{t-1})$ being Poisson(λ_{t-1}). It follows that Assumption 1 in the paper by Meitz and Saikonnen (2008) is fulfilled. Therefore, Prop. 1 of Meitz and Saikonnen (2008) shows that $\{Y_t^m, U_t, \lambda_t^m)\}$ inherits the geometric ergodicity of $\{\lambda_t^m\}$. Moreover, since we can take $\mathrm{E}_\mu[1 + (d + \lambda_0^m + Y_0^m + \varepsilon_{1,m})^k] < \infty$, where μ is the distribution of the initial value $(Y_0^m, U_1, \lambda_0^m)$, and since for some constant $C(k)$

$$\begin{aligned}\mathrm{E}[1 + (\lambda_{t-1}^m)^k + (Y_{t-1}^m)^k + \varepsilon_{t,m}^k | \lambda_{t-1}^m = \lambda] &= C(k) + \lambda^k + \mathrm{E}[Y_{t-1}^k | \lambda_{t-1} = \lambda] \\ &= C(k) + \lambda^k + \lambda^k + \sum_{i=1}^{k-1} c_i \lambda^i \\ &\leqslant C'(k)(1 + \lambda^k),\end{aligned}$$

it follows from Thm. 2 of Meitz and Saikonnen (2008)–see also Prop.2– that $\{(Y_t^m, U_t, \lambda_t^m)\}$ is $V_{(Y,U,\lambda)}$- geometrically ergodic with $V_{(Y,U,\lambda)}(Y, U, \lambda) = 1 + Y^k + U^k + \lambda^k$.

Proof of Lemma 2.1

It follows from the defining equations (1) and (2)

$$\lambda_t^m - \lambda_t = a\left(\lambda_{t-1}^m - \lambda_{t-1}\right) + b\left(Y_{t-1}^m - Y_{t-1}\right) + \varepsilon_{t,m}, \tag{A-2}$$

and by taking conditional expectation and using the properties of the Poisson process we obtain that

$$\mathrm{E}\left(\lambda_t^m - \lambda_t\right) = (a+b)\,\mathrm{E}\left(\lambda_{t-1}^m - \lambda_{t-1}\right) + \mathrm{E}\left(\varepsilon_{t,m}\right) = \sum_{i=0}^{t-1} (a+b)^i\,\mathrm{E}\left(\varepsilon_{t-i,m}\right).$$

Since $(a+b) < 1$ and $|\mathrm{E}\left(\varepsilon_{t,m}\right)| \leqslant c_m$, with $c_m \to 0$ as $m \to \infty$,

$$\left|\mathrm{E}\left(\lambda_t^m - \lambda_t\right)\right| \leqslant \frac{c_m}{1-(a+b)} := \delta_{1,m},$$

which proves the first assertion.

Next consider the second statement. By using (A-2) again,

$$\begin{aligned}\mathrm{E}\left(\lambda_t^m - \lambda_t\right)^2 &= a^2\mathrm{E}\left(\lambda_{t-1}^m - \lambda_{t-1}\right)^2 + b^2\mathrm{E}\left(Y_{t-1}^m - Y_{t-1}\right)^2 \\ &+ 2ab\mathrm{E}\left(Y_{t-1}^m - Y_{t-1}\right)\left(\lambda_{t-1}^m - \lambda_{t-1}\right) + \mathrm{E}\left(\varepsilon_{t,m}^2\right) \\ &+ 2a\mathrm{E}\left[\left(\lambda_{t-1}^m - \lambda_{t-1}\right)\varepsilon_{t,m}\right] + 2b\mathrm{E}\left[\left(Y_{t-1}^m - Y_{t-1}\right)\varepsilon_{t,m}\right].\end{aligned}$$

However, for $\lambda_t^m \geqslant \lambda_t$

$$\begin{aligned}\mathrm{E}\left(\left(Y_t^m - Y_t\right)\left(\lambda_t^m - \lambda_t\right)\right) &= \mathrm{E}\left[\mathrm{E}\left(\left(Y_t^m - Y_t\right)\left(\lambda_t^m - \lambda_t\right)\right)|\mathcal{F}_{t-1}\right] \\ &= \mathrm{E}\left[\left(\lambda_t^m - \lambda_t\right)\left(\mathrm{E}\left(N_t[\lambda_t, \lambda_t^m]\right)\right)\right] = \mathrm{E}\left(\lambda_t^m - \lambda_t\right)^2,\end{aligned}$$

where $\mathcal{F}_{t-1}$ is the σ-algebra generated by $\{(U_{k+1}, N_k, k \leqslant t-1\}$ and $N_t[\lambda_t, \lambda_t^m]$ is equal to the number of events between λ_t and λ_t^m for the unit intensity Poisson process N_t. (If $\lambda_t^m < \lambda_t$, we work along the same lines). By using again the properties of the Poisson process, we find that

$$\mathrm{E}\left(Y_t^m - Y_t\right)^2 \leqslant \mathrm{E}\left(\lambda_t^m - \lambda_t\right)^2 + 2\left|\mathrm{E}\left(\lambda_t^m - \lambda_t\right)\right| \leqslant \mathrm{E}\left(\lambda_t^m - \lambda_t\right)^2 + 2\delta_{1,m}. \tag{A-3}$$

Finally, with K a positive constant,

$$\mathrm{E}\left(\varepsilon_{t,m}^2\right) + 2a\mathrm{E}\left(\left(\lambda_{t-1}^m - \lambda_{t-1}\right)\varepsilon_{t,m}\right) + 2b\mathrm{E}\left(\left(Y_{t-1}^m - Y_{t-1}\right)\varepsilon_{t,m}\right) \leqslant Kc_m^2.$$

Therefore, by simple recursion

$$\mathrm{E}\left(\lambda_t^m - \lambda_t\right)^2 \leqslant (a+b)^2\,\mathrm{E}\left(\lambda_{t-1}^m - \lambda_{t-1}\right)^2 + \kappa c_m^2 + b^2 2\delta_{1,m} \leqslant \delta_{2,m},$$

where $\delta_{2,m} \to 0$ as $m \to \infty$. This establishes the second assertion and hence the last statement of the Lemma. As to third statement this follows by (A-3) and the second assertion of the Lemma. □

Proof of Lemma 3.2

Equation (9) shows that the score for the perturbed model is given by $S_n^m(\theta) = \sum_{t=1}^n \partial l_t^m(\theta)/\partial\theta$, with martingale difference terms defined by

$$\frac{\partial l_t^m}{\partial\theta} = \left(\frac{Y_t^m}{\lambda_t^m} - 1\right)\frac{\partial\lambda_t^m}{\partial\theta} := Z_t^m\frac{\partial\lambda_t^m}{\partial\theta}$$

where $\partial\lambda_t^m/\partial\theta$ is defined analogous to (10). It follows that at $\theta = \theta_0$, $\mathrm{E}\left(\partial l_t^m/\partial\theta|\mathcal{F}_{t-1}\right) = 0$ and $\mathrm{E}\left((Z_t^m)^2|\mathcal{F}_{t-1}\right) = 1/\lambda_t^m$ where $\mathcal{F}_{t-1}$ is the σ-algebra generated by $\{U_{k+1}, N_k, k \leqslant t-1\}$. Furthermore, from (3) and (10),

$$\frac{\partial\lambda_t^m}{\partial a} = \sum_{i=0}^{t-1} a_0^i\lambda_{t-1-i}^m, \quad \frac{\partial\lambda_t^m}{\partial d} = \frac{1-a_0^t}{1-a_0} \quad \text{and} \quad \frac{\partial\lambda_t^m}{\partial b} = \sum_{i=0}^{t-1} b_0^i Y_{t-1-i}^m.$$

Observe that, as $a_0, b_0 < 1$, $\mathrm{E}\left(Y_t^m\right)^2 < \infty$, and $\mathrm{E}\left(\lambda_t^m\right)^2 < \infty$, then $\left(\partial\lambda_t^m/\partial d\right)^2$, $\mathrm{E}\left(\partial\lambda_t^m/\partial a\right)^2$ and $\mathrm{E}\left(\partial\lambda_t^m/\partial b\right)^2$ are all finite. Also $1/\lambda_t^m \leqslant 1/\left(d_0 - \eta_1\right)$ for any small $\eta_1 \geqslant 0$, where for m large enough,

$$\varepsilon_{t,m} = c_m 1\left(Y_{t-1}^m = 1\right) U_t \in [-\eta_1, \eta_1].$$

From Hölders inequality we conclude that $\mathrm{E}\left\|\partial l_t^m/\partial\theta\right\| < \infty$. Thus $\partial l_t^m/\partial\theta$ is a martingale difference sequence with respect to $\mathcal{F}_t$ and an application of the CLT (Hall and Heyde, 1980, Cor. 3.1), gives $n^{-1/2}S_n^m$ is asymptotically Gaussian with covariance given by the limit,

$$\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(Z_t^2\left(\frac{\partial\lambda_t^m}{\partial\theta}\right)\left(\frac{\partial\lambda_t^m}{\partial\theta}\right)'|\mathcal{F}_{t-1}\right) \xrightarrow{P} G^m,$$

by the LLN for geometrically ergodic process in Jensen and Rahbek (2007). That the conditional Lindeberg's condition holds follows by noting

$$\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(\left\|\partial l_t^m/\partial\theta\right\|^2 I\left(\left\|\partial l_t^m/\partial\theta\right\| > \sqrt{n}\delta\right)|\mathcal{F}_{t-1}\right) \leqslant \frac{1}{n^2\delta^2}\sum_{t=1}^{n}\mathrm{E}\left(\left\|\partial l_t^m/\partial\theta\right\|^4|\mathcal{F}_{t-1}\right) \to 0,$$

since $\mathrm{E}\left\|\partial l_t^m/\partial\theta\right\|^4 < \infty$. This proves the first assertion of the Lemma. The second assertion follows by Lemma 3.1.

We consider the last conclusion of the Lemma. Define Z_t in a analogous way as Z_t^m. Then

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(S_n^m - S_n\right) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(\frac{\partial l_t^m}{\partial\theta} - \frac{\partial l_t}{\partial\theta}\right) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(Z_t^m\frac{\partial\lambda_t^m}{\partial\theta} - Z_t\frac{\partial\lambda_t}{\partial\theta}\right)\\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[Z_t^m\left(\frac{\partial\lambda_t^m}{\partial\theta} - \frac{\partial\lambda_t}{\partial\theta}\right) + \left(Z_t^m - Z_t\right)\frac{\partial\lambda_t}{\partial\theta}\right].\end{aligned}$$

But for the first summand

$$\begin{aligned}\mathrm{P}\left(\left\|\sum_{t=1}^{n}Z_t^m\left(\frac{\partial\lambda_t^m}{\partial\theta} - \frac{\partial\lambda_t}{\partial\theta}\right)\right\| > \delta\sqrt{n}\right) &\leqslant \mathrm{P}\left(\gamma_m\left\|\sum_{t=1}^{n}Z_t^m\right\| > \delta\sqrt{n}\right)\\ &\leqslant \frac{\gamma_{a,m}^2}{\delta^2 n}\sum_{t=1}^{n}\mathrm{E}\left\|Z_t^m\right\|^2 \leqslant C\gamma_{a,m}^2 \to 0,\end{aligned}$$

as $m \to \infty$. For the second summand, the same arguments apply because $\mathrm{E}\left\|\partial\lambda_t/\partial\theta\right\|^2 < \infty$. In addition, for

$$Z_t^m - Z_t = \frac{\lambda_t\left(Y_t^m - Y_t\right) - Y_t\left(\lambda_t^m - \lambda_t\right)}{\lambda_t\lambda_t^m},$$

we have $\|Z_t^m - Z_t\| < \gamma_{b,m} \to 0$ as $m \to \infty$. To see this, observe

$$\mathrm{E}\left|\frac{\left(Y_t^m - Y_t\right)}{\lambda_t^m}\right| \leqslant \mathrm{E}\left|Y_t^m - Y_t\right| / \left(d_0 - \eta_1\right) \leqslant C\delta_{1,m} \quad \text{and} \quad \mathrm{E}\left|\frac{Y_t\left(\lambda_t^m - \lambda_t\right)}{\lambda_t\lambda_t^m}\right| \leqslant \frac{\delta_m \mathrm{E}\left|Y_t\right|}{d_0\left(d_0 - \eta_1\right)} \leqslant C\delta_m$$

using Lemma 2.1 and the fact that $\mathrm{E}|Y_t| < \infty$.

References

Aue, A., I. Berkes, and L. Horváth (2006). Strong approximation for the sums of squares of augmented GARCH sequences. *Bernoulli 12*, 583–608.

Benjamin, M. A., R. A. Rigby, and D. M. Stasinopoulos (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association 98*, 214–223.

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

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Data Analysis and Theory* (2nd ed.). New York: Springer.

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

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory 18*, 1–39.

Chan, K. S. (1990). *Deterministic stability, stochastic stability and ergodicity*, Appendix 1 of H. Tong (1990), Non–Linear Time Series: A Dynamical System Approach. Clarendon Press, Oxford.

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

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 74*, 33–43.

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

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer.

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

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

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

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

Hall, P. and C. C. Heyde (1980). *Martingale Limit Theory and its Applications*. New York: Academic Press.

Jensen, S. T. and A. Rahbek (2004). Asymptotic inference for nonstationary GARCH. *Econometric Theory 20*, 1203–1226.

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

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

Kedem, B. and K. Fokianos (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.

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

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

Mikosch, T. and D. Straumann (2006). Stable limits of martingale transforms with application to the estimation of GARCH parameters. *The Annals of Statistics 34*, 493–522.

Rydberg, T. H. and N. Shephard (2000). A modeling framework for the prices and times of trades on the New York stock exchange. In W. J. Fitzgerlad, R. L. Smith, A. T. Walden, and P. C. Young (Eds.), *Nonlinear and Nonstationary Signal Processing*, pp. 217–246. Cambridge: Isaac Newton Institute and Cambridge University Press.

Streett, S. (2000). *Some observation driven models for time series of counts*. Ph. D. thesis, Colorado State University, Department of Statistics.

Tong, H. (1990). *Nonlinear Time Series:A Dynamical System Approach*. New York: Oxford University Press.

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