Statistica Sinica **26** (2016), 1673-1707
doi:http://dx.doi.org/10.5705/ss.2014.145t

THEORY AND INFERENCE FOR A CLASS OF NONLINEAR MODELS WITH APPLICATION TO TIME SERIES OF COUNTS

Richard A. Davis and Heng Liu

Columbia University and Google Inc.

Abstract: This paper studies theory and inference related to a class of time series models that incorporates nonlinear dynamics. It is assumed that the observations follow a one-parameter exponential family of distributions given an accompanying process that evolves as a function of lagged observations. We employ an iterated random function approach and a special coupling technique to show that, under suitable conditions on the parameter space, the conditional mean process is a geometric moment contracting Markov chain and that the observation process is absolutely regular with geometrically decaying coefficients. Asymptotic theory of the maximum likelihood estimates of the parameters is established under some mild assumptions. These models are applied to two examples; the first is the number of transactions per minute of Ericsson stock and the second is related to return times of extreme events of Goldman Sachs Group stock.

Key words and phrases: Absolute regularity, ergodicity, geometric moment contraction, iterated random functions, one-parameter exponential family, time series of counts.

1. Introduction

With a surge in the range of applications from economics, finance, environmental science, social science and epidemiology, there has been renewed interest in developing models for time series of counts. Many of these models assume that the observations follow a Poisson distribution conditioned on an accompanying intensity process that drives the dynamics of the models, e.g., Davis, Dunsmuir, and Wang (2000), Davis, Dunsmuir, and Streett (2003), Fokianos, Rahbek, and Tjøstheim (2009), Neumann (2011), Streett (2000) and Doukhan, Fokianos, and Tjøstheim (2012). According to whether the evolution of the intensity process depends on the observations or solely on an external process, Cox (1981) classified the models into observation-driven and parameter-driven. This paper focuses on the theory and inference for a particular class of observation-driven models.

Many of the models proposed in the literature, such as the Poisson integer-valued GARCH (INGARCH), are special cases of our model. For an INGARCH,

the observations $\{Y_t\}$ given the intensity process $\{\lambda_t\}$ follow a Poisson distribution and λ_t is a linear combination of its lagged values and lagged Y_t. The model is capable of capturing positive temporal correlation in the observations and it is relatively easy to fit via maximum likelihood. Ferland, Latour, and Oraichi (2006) showed second moment stationarity through a sequence of approximating processes and Fokianos, Rahbek, and Tjøstheim (2009) established the consistency and asymptotic normality of the MLE by introducing a perturbed model. However, these results rely heavily on the Poisson assumption and the GARCH-like dynamics of λ_t. Later Zhu (2010) and Zhu (2012b) considered negative binomial and zero-inflated Poisson-based models. But the authors only established conditions under which the processes are moment stationary, which is not sufficient to study such inference properties as the asymptotic behavior of the maximum likelihood estimates. Neumann (2011) relaxed the linear assumption on the Poisson INGARCH to a general contracting evolution rule and proved the absolute regularity for this Poisson count process, and Doukhan, Fokianos, and Tjøstheim (2012) showed the existence of moments under similar conditions by utilizing the concept of weak dependence. More recently, Blasques, Koopman, and Lucas (2012) considered a class of generalized autoregressive score processes which includes the integer-valued GARCH as a special case and used the Dudley entropy integral to obtain a wider non-degenerate parameter region that guarantees the stationarity and ergodicity of the processes. Zhu (2012a) considered the INGARCH with generalized Poisson as the conditional distribution.

In our study the conditional distribution of the observation Y_t given the past is assumed to follow a one-parameter exponential family. The temporal dependence in the model is defined through recursions relating the conditional mean process X_t with its lagged values and lagged observations. Theory from iterated random functions (IRF), see e.g., Diaconis and Freedman (1999) and Wu and Shao (2004), is utilized to establish some key stability properties, such as existence of a stationary and mixing solution. This theory allows us to consider both linear and nonlinear dynamic models as well as inference questions. In particular, the asymptotic normality of the maximum likelihood estimates can be established. The nonlinear dynamic models are also investigated in a simulation study and both linear and nonlinear models are applied to two datasets.

The organization of the paper is as follows. Section 2 formulates the model and establishes stability properties. The maximum likelihood estimates of the parameters and the relevant asymptotic theory are derived in Section 3. Examples of both linear and nonlinear dynamic models are considered in Section 4. Numerical results, including a simulation study and two data applications are given in Section 5, where the models are applied to the number of transactions per minute of Ericsson stock and to the return times of extreme events

of Goldman Sachs Group (GS) stock. Some diagnostic tools for assessing and comparing model performance are also given in Section 5. Appendix A reviews some standard properties of the one-parameter exponential family, Appendix B summarizes application to the linear dynamic models, and Appendix C contains the proofs of the key results in Sections 2−4.

2. Model Formulation and Stability Properties

2.1. One-parameter exponential family

A random variable Y is said to follow a distribution of the one-parameter exponential family if its probability density function with respect to some σ-finite measure μ is given by

$$p(y|\eta) = \exp\{\eta y - A(\eta)\}h(y), \quad y \geq 0, \tag{2.1}$$

where η is the natural parameter, and $A(\eta)$ and $h(y)$ are known functions. If $B(\eta) = A'(\eta)$, then it is known that $\mathrm{E}Y = B(\eta)$ and $\mathrm{Var}(Y) = B'(\eta)$. The derivative of $A(\eta)$ generally exists for the exponential family, see e.g., Lehmann and Casella (1998). Since $B'(\eta) = \mathrm{Var}(Y) > 0$, so $B(\eta)$ is strictly increasing, which establishes a one-to-one association between the values of η and $B(\eta)$. Moreover, because we assume that the support of Y is non-negative throughout this paper, $B(\eta) = \mathrm{E}Y > 0$, which implies that $A(\eta)$ is strictly increasing. Other properties of this family of distributions are presented in Appendix A.

Many familiar distributions belong to this family, including Poisson, negative binomial, Bernoulli, exponential, etc. If the shape parameter is fixed, then the gamma distribution is also a member of this family. While we restrict consideration to the univariate case, extensions to the multi-parameter exponential family is a topic of future research.

2.2. Model formulation

Set $\mathcal{F}_0 = \sigma\{\eta_1\}$, where η_1 is a natural parameter of (2.1), assumed fixed for the moment. Let $Y_1, Y_2, \ldots$ be observations from a model that is defined recursively as

$$Y_t|\mathcal{F}_{t-1} \sim p(y|\eta_t), \quad X_t = g_\theta(X_{t-1}, Y_{t-1}) \tag{2.2}$$

for all $t \geq 1$, where $p(y|\eta_t)$ is defined in (2.1), $\mathcal{F}_t = \sigma\{\eta_1, Y_1, \ldots, Y_t\}$, and X_t is the conditional mean process, $X_t = B(\eta_t) = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$. Here $g_\theta(x, y)$ is a non-negative bivariate function defined on $[0, \infty) \times [0, \infty)$ when Y_t has a continuous conditional distribution, or on $[0, \infty) \times \mathbb{N}_0$, where $\mathbb{N}_0 = \{0, 1, \ldots\}$, when Y_t only takes non-negative integers. Throughout, we assume that the function g_θ satisfies a contraction condition: for any $x, x' \geq 0$, and $y, y' \in [0, \infty)$ or $\mathbb{N}_0$,

$$|g_\theta(x, y) - g_\theta(x', y')| \leq a|x - x'| + b|y - y'|, \tag{2.3}$$

where a and b are non-negative constants with $a + b < 1$. Here (2.3) implies

$$g_\theta(x, y) \le g_\theta(0, 0) + ax + by, \quad \text{for any} \quad x, y \ge 0. \tag{2.4}$$

Model (2.2) with the function g_θ satisfying (2.3) includes the Poisson INGARCH model (see Example 1 in Section 4.1) and the exponential autoregressive model (4.11) as special cases under some restrictions on the parameter space. The generalized linear autoregressive moving average model (GLARMA) (see Davis, Dunsmuir, and Streett (2003)) also belongs to this class, although the contraction condition is not necessarily satisfied. Only under very simple model specifications have the stability properties of GLARMA been established and the relevant work is still ongoing. The primary focus of this paper is on the conditional mean process $\{X_t\}$ that, by the dynamics described in (2.2), is easily seen to be a time-homogeneous Markov chain. The observation process $\{Y_t\}$ is not a Markov chain itself.

2.3. Strict stationarity

The iterated random function approach (see e.g., Diaconis and Freedman (1999) and Wu and Shao (2004)) provides a useful tool when investigating the stability properties of Markov chains, and is particularly instrumental in our research. In the definition of iterated random functions (IRF), the state space $(\mathcal{W}, \rho)$ is assumed to be a complete and separable metric space. Then a sequence of *iterated random functions* $\{f_{U_t}\}$ is defined through

$$W_t = f_{U_t}(W_{t-1}), \quad t \in \mathbb{N},$$

where $\{U_t\}_{t\ge1}$ take values in another measurable space Ψ and are independently distributed with identical marginal distribution, and W_0 is independent of $\{U_t\}_{t\ge1}$.

In working with iterated random functions, Wu and Shao (2004) introduced the idea of geometric moment contraction (GMC), which is useful for deriving further properties of IRF. Our research also relies heavily on GMC. Suppose there exists a stationary solution to the Markov chain $\{W_t\}$, denoted by ϖ, let $W_0, W_0' \sim \varpi$ be independent of each other and of $\{U_t\}_{t\ge1}$, and define $W_t(w) = f_{U_t} \circ f_{U_{t-1}} \circ \ldots \circ f_{U_1}(w)$. Then $\{W_t\}$ is said to be *geometric moment contracting* if there exist an $\alpha > 0$, a $C = C(\alpha) > 0$, and an $r = r(\alpha) \in (0, 1)$ such that, for all $t \in \mathbb{N}$,

$$\mathrm{E}\{\rho^\alpha(W_t(W_0), W_t(W_0'))\} \le Cr^t.$$

The conditional mean process $\{X_t\}$ specified in (2.2) can be embedded into the framework of IRF and shown to be GMC.

In this section and the next we use g to represent the function g_θ in (2.2) evaluated at the true parameter. For any $u \in (0,1)$, the random function $f_u(x)$ is

$$f_u(x) := g\big(x, F_x^{-1}(u)\big), \tag{2.5}$$

where F_x is the cumulative distribution function of $p(y|\eta)$ in (2.1) with $x = B(\eta)$, and its inverse $F_x^{-1}(u) := \inf\{t \geq 0 : F_x(t) \geq u\}$ for $u \in [0,1]$. Let $\{U_t\}$ be a sequence of independent and identically distributed (iid) uniform $(0,1)$ random variables, then the Markov chain $\{X_t\}$ defined in (2.2) starting from $X_0 = x$ can be represented as the so-called forward process $X_t(x) = (f_{U_t} \circ f_{U_{t-1}} \circ \ldots \circ f_{U_1})(x)$. The corresponding backward process is $Z_t(x) = (f_{U_1} \circ f_{U_2} \circ \ldots \circ f_{U_t})(x)$, which has the same distribution as $X_t(x)$ for any t.

Proposition 1. *If* (2.2) *holds, and* g *satisfies* (2.3)*, then*

1. *there exists a random variable* Z_∞ *such that, for all* $x \in S$, $Z_n(x) \to Z_\infty$ *almost surely,* Z_∞ *does not depend on* x *and has distribution* π*, the stationary distribution of* $\{X_t\}$*;*
2. *the Markov chain* $\{X_t, t \geq 1\}$ *is geometric moment contracting with* π *as its unique stationary distribution, and* $E_\pi X_1 < \infty$*;*
3. *if* $\{X_t, t \geq 1\}$ *starts from* π*,* $X_1 \sim \pi$*, then* $\{Y_t, t \geq 1\}$ *is a stationary time series.*

Proposition 1 implies that, starting from any state x, the limiting distribution of the Markov chain $X_n(x)$ exists and the n-step transition probability measure $P^n(x, \cdot)$ converges weakly to π as $n \to \infty$.

2.4. Ergodicity

In this section we investigate the stability properties, including ergodicity and mixing, under (2.2). Under the conditions of Proposition 1, the process $\{(X_t, Y_t)\}$ is strictly stationary, so we can extend it to be indexed by all the integers. The following proposition establishes ergodicity and absolute regularity when Y_t is discrete.

Proposition 2. *Assume* (2.2) *with the support of* Y_t *a subset of* $\mathbb{N}_0 = \{0, 1, \ldots, \}$*, and that* g *satisfies* (2.3)*. Then*

1. *there exists a measurable function* $g_\infty : \mathbb{N}_0^\infty = \{(n_1, n_2, \ldots), n_i \in \mathbb{N}_0, i = 1, 2, \ldots\} \longrightarrow [0, \infty)$ *such that* $X_t = g_\infty(Y_{t-1}, Y_{t-2}, \ldots)$ *almost surely;*
2. *the count process* $\{Y_t\}$ *is absolutely regular with coefficients satisfying*

$$\beta(n) \leq \frac{(a+b)^n}{1-(a+b)},$$

and hence $\{(X_t, Y_t)\}$ *is ergodic.*

3. Likelihood Inference

In this section, we consider maximum likelihood estimates of the parameters and study their asymptotic behavior. Denote the $d-$dimensional parameter vector by $\theta \in \mathbb{R}^d$, and the true parameter vector by $\theta_0 = (\theta_1^0, \ldots, \theta_d^0)^T$. Then the likelihood function of (2.2), conditioned on η_1 and based on the observations $Y_1, \ldots, Y_n$, is

$$L(\theta | Y_1, \ldots, Y_n, \eta_1) = \prod_{t=1}^{n} \exp\{\eta_t(\theta) Y_t - A(\eta_t(\theta))\} h(Y_t),$$

where $\eta_t(\theta) = B^{-1}(X_t(\theta))$ is updated through the iterations $X_t = g_\theta(X_{t-1}, Y_{t-1})$. The log-likelihood function is, up to a constant independent of θ,

$$l(\theta) = \sum_{t=1}^{n} l_t(\theta) = \sum_{t=1}^{n} \{\eta_t(\theta) Y_t - A(\eta_t(\theta))\}, \tag{3.1}$$

with score function

$$S_n(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n} \{Y_t - B(\eta_t(\theta))\} \frac{\partial \eta_t(\theta)}{\partial \theta}. \tag{3.2}$$

The maximum likelihood estimator $\hat{\theta}_n$ is a solution to $S_n(\theta) = 0$. Let P_{θ_0} be the probability measure under the true parameter θ_0 and, unless otherwise indicated, $\mathrm{E}[\cdot]$ is taken under θ_0. Recall that $X_t = g_\infty^\theta(Y_{t-1}, Y_{t-2}, \ldots)$ according to part (a) of Propositions 2. We derive the asymptotic properties of the maximum likelihood estimator $\hat{\theta}_n$ based on a set of regularity conditions

(A0) θ_0 is an interior point in the compact parameter space $\Theta \in \mathbb{R}^d$.

(A1) For any $\theta \in \Theta$, $g_\infty^\theta \geq x_\theta^* \in \mathcal{R}(B)$, where $\mathcal{R}(B)$ is the range of $B(\eta)$. Moreover $x_\theta^* \geq x^* \in \mathcal{R}(B)$ for all θ.

(A2) For any $\mathbf{y} \in [0, \infty)^\infty$ or $\mathbb{N}_0^\infty$, the mapping $\theta \mapsto g_\infty^\theta(\mathbf{y})$ is continuous.

(A3) $g(x, y)$ is increasing in (x, y) if Y_t given $\mathcal{F}_{t-1}$ has a continuous distribution.

(A4) $\mathrm{E}\{Y_1 \sup_{\theta \in \Theta} B^{-1}(g_\infty^\theta(Y_0, Y_{-1}, \ldots))\} < \infty$.

(A5) If there exists a $t \geq 1$ such that $X_t(\theta) = X_t(\theta_0)$, P_{θ_0}-a.s., then $\theta = \theta_0$.

(A6) The mapping $\theta \mapsto g_\infty^\theta$ is twice continuously differentiable.

(A7) $\mathrm{E}\{B'(\eta_1(\theta_0))(\partial \eta_1(\theta)/\partial \theta_i)^2|_{\theta=\theta_0}\} < \infty$, for $i = 1, \ldots, d$.

Strong consistency of the estimates is derived from a lemma that is adapted from Lemma 3.11 in Pfanzagl (1969).

Lemma 1. *Assume that $\Theta \subset \mathbb{R}^d$ is a compact set, and that $(\Omega, \mathcal{F}, P)$ is a probability space. Let $\{f_\theta : \mathbb{R}^\infty \mapsto [-\infty, \infty], \theta \in \Theta\}$ be a family of Borel measurable functions such that $\theta \mapsto f_\theta(\mathbf{x})$ is upper-semicontinuous for all $\mathbf{x} \in \mathbb{R}^\infty$, $\sup_{\theta \in C} f_\theta(\mathbf{x})$ is Borel measurable for any compact set $C \subset \Theta$, and $E\{\sup_{\theta \in \Theta} f_\theta(X)\} < \infty$ for some random variable X defined on $(\Omega, \mathcal{F}, P)$. Then*

1. *$\theta \mapsto E[f_\theta(X)]$ is upper-semicontinuous;*
2. *if $\{X_t : \Omega \mapsto \mathbb{R}^\infty, t \in \mathbb{Z}\}$ is an ergodic stationary process defined on $(\Omega, \mathcal{F}, P)$ and, for all t, X_t has the same distribution as X, then*

$$\limsup_{n\to\infty} \sup_{\theta \in C} \frac{1}{n} \sum_{i=1}^{n} f_\theta(X_i) \le \sup_{\theta \in C} E\{f_\theta(X_1)\}, \quad a.s.\text{-}P,$$

for any compact set C.

Theorem 1. *If (2.2) holds with g satisfying (2.3), and (A0)−(A5) hold, then $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$, as $n \to \infty$.*

The asymptotic distribution of the MLE and its proof is similar to that in Davis, Dunsmuir, and Streett (2003). Unless otherwise indicated, $\eta_t = \eta_t(\theta_0)$ and $\dot{\eta}_t = (\partial \eta_t / \partial \theta)|_{\theta=\theta_0}$.

Theorem 2. *If (2.2) holds with g satisfying (2.3), and (A0)−(A7) hold, then $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega^{-1})$, as $n \to \infty$, where $\Omega = E\{B'(\eta_t)\dot{\eta}_t \dot{\eta}_t^T\}$.*

In practice, the population quantities in Ω can be replaced by their estimated counterparts. Examples are given in specific models.

4. Examples

4.1. Linear dynamic models

The conditional mean process $\{X_t\}$ in these models has GARCH-like dynamics, and specifically

$$Y_t | \mathcal{F}_{t-1} \sim p(y|\eta_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1}, \tag{4.1}$$

where $X_t = B(\eta_t) = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$, and $\delta > 0, \alpha, \beta \ge 0$ are parameters. Model (4.1) is a special case of (2.2) with

$$g_\theta(x, y) = \delta + \alpha x + \beta y, \tag{4.2}$$

where $\theta = (\delta, \alpha, \beta)^T$ and (2.3) corresponds to $\alpha + \beta < 1$. Under this condition, $\{Y_t\}$ can be represented as a causal ARMA(1,1) process. To see this, if $d_t = Y_t - X_t$, then it follows from $\mathrm{E}(d_t|\mathcal{F}_{t-1}) = 0$ that $\{d_t, t \in \mathbb{Z}\}$ is a martingale difference sequence. Therefore (4.1) can be written as

$$Y_t - (\alpha + \beta) Y_{t-1} = \delta + d_t - \alpha d_{t-1}. \tag{4.3}$$

Let $\gamma_Y(h)$ denote the auto-covariance function of $\{Y_t\}$. If $\gamma_Y(0) < \infty$, then $\gamma_Y(h) = (\alpha+\beta)^{h-1}\gamma_Y(1)$, for $h \geq 1$, see for example Brockwell and Davis (1991). A direct application of Propositions 1 and 2 gives the stability properties of (4.1).

Proposition 3. *If* (4.1) *holds with* $\alpha+\beta<1$, *then* $\{X_t, t\geq 1\}$ *has a unique stationary distribution* π, *and* $\{(X_t,Y_t), t\geq 1\}$ *is ergodic if* $X_1 \sim \pi$.

If $\theta_0 = (\delta_0, \alpha_0, \beta_0)^T$ denotes the true parameter vector, then the log-likelihood function $l(\theta)$ and the score function $S_n(\theta)$ of (4.1) are given by (3.1) and (3.2), respectively, where $\partial\eta_t(\theta)/\partial\theta = (\partial\eta_t/\partial\delta, \partial\eta_t/\partial\alpha, \partial\eta_t/\partial\beta)^T$ is determined recursively by

$$\frac{\partial\eta_t}{\partial\theta} = \begin{pmatrix} 1 \\ B(\eta_{t-1}) \\ Y_{t-1} \end{pmatrix} / B'(\eta_t) + \alpha \frac{B'(\eta_{t-1})}{B'(\eta_t)} \frac{\partial\eta_{t-1}}{\partial\theta}. \tag{4.4}$$

The maximum likelihood estimator $\hat{\theta}_n$ is a solution of $S_n(\theta) = 0$. In order to apply Theorem 2 when investigating the asymptotic behavior of the MLE, we need to impose regularity conditions

(L0) θ_0 lies in a compact neighborhood $\Theta \in \mathbb{R}^3_+$ of θ_0, where $\Theta = \{\theta = (\delta,\alpha,\beta)^T \in \mathbb{R}^3_+ : 0<\delta_L \leq \delta \leq \delta_U, \epsilon \leq \alpha+\beta \leq 1-\epsilon\}$ for some $\epsilon > 0$.

(L1) $\mathrm{E}\{Y_1 \sup_{\theta\in\Theta} B^{-1}(\delta/(1-\alpha) + \beta\sum_{k=0}^{\infty}\alpha^k Y_{-k})\} < \infty$.

(L2) $\mathrm{E}\{B'(\eta_1(\theta_0))(\partial\eta_1(\theta)/\partial\theta_i)^2|_{\theta=\theta_0}\} < \infty$, for $i = 1,2,3$.

Theorem 3. *If* (4.1) *and* (L0)−(L2) *hold, the maximum likelihood estimator* $\hat{\theta}_n$ *is strongly consistent and*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega^{-1}), \quad \textit{as} \;\; n\to\infty,$$

where $\Omega = E\{B'(\eta_t)\dot{\eta}_t\dot{\eta}_t^T\}$, $\eta_t = \eta_t(\theta_0)$, *and* $\dot{\eta}_t = \frac{\partial\eta_t}{\partial\theta}|_{\theta=\theta_0}$.

In practice, it can be difficult to verify (L1) and (L2), so we provide some alternative sufficient conditions for them. Proofs can be found in Appendix B.

Remark 1. A sufficient condition for (L1) is

$$\mathrm{E}\Big\{Y_1 B^{-1}\Big(\frac{\delta_U}{\epsilon} + \sum_{k=1}^{\infty}(1-\epsilon)^k Y_{1-k}\Big)\Big\} < \infty,$$

provided $\delta_U/\epsilon + \sum_{k=1}^{\infty}(1-\epsilon)^k Y_{1-k}$ is in the range of $B(\eta)$.

Remark 2. If $A''(\eta_t) \geq \underline{c}$ for some $\underline{c} > 0$ (true, for example, when $A''(\eta)$ is increasing and $A''(B^{-1}(\delta_L)) > 0$), then a sufficient condition for (L2) is $\gamma_Y(0) < \infty$.

Example 1. The Poisson INGARCH$(1,1)$ model is

$$Y_t|\mathcal{F}_{t-1} \sim \text{Pois}(\lambda_t), \quad \lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1}, \tag{4.5}$$

where $\delta > 0, \alpha, \beta \geq 0$ are parameters. If $\alpha + \beta < 1$, then $\{\lambda_t\}$ has a unique stationary distribution π; moreover if $\lambda_1 \sim \pi$, then $\{(Y_t, \lambda_t), t \geq 1\}$ is ergodic. In addition, under the assumption (L0), the maximum likelihood estimator $\hat{\theta}_n$ is consistent and asymptotically normal.

The iterated random function approach can be used to study the properties of INGARCH models with higher orders. A Poisson INGARCH(p,q) model takes the form

$$Y_t|\mathcal{F}_{t-1} \sim \text{Pois}(\lambda_t), \quad \lambda_t = \delta + \sum_{i=1}^{p} \alpha_i \lambda_{t-i} + \sum_{j=1}^{q} \beta_j Y_{t-j}, \tag{4.6}$$

where $\delta > 0, \alpha_i, \beta_j \geq 0, i = 1, \ldots, p;\ j = 1, \ldots, q$. Applying similar ideas as in the INGARCH$(1,1)$ case, we have the following.

Proposition 4. *Consider the INGARCH(p,q) model* (4.6). *If* $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$, *then* $\{\lambda_t\}$ *is geometric moment contracting and has a unique stationary distribution.*

The proof can be found in Appendix B.3.

Example 2. The negative binomial INGARCH$(1,1)$ model (NB-INGARCH) is

$$Y_t|\mathcal{F}_{t-1} \sim \text{NB}(r, p_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1}, \tag{4.7}$$

where $X_t = r(1-p_t)/p_t$, $\delta > 0, \alpha, \beta \geq 0$ are parameters, and the notation $Y \sim \text{NB}(r,p)$ means

$$P(Y = k) = \binom{k+r-1}{r-1}(1-p)^k p^r, \qquad k = 0, 1, 2, \ldots.$$

When $r = 1$, the conditional distribution of Y_t is geometric with probability of success p_t, in which case (4.7) reduces to a geometric INGARCH model.

By virtue of Proposition 3, if $\alpha + \beta < 1$, then $\{X_t, t \geq 1\}$ is a geometric moment contracting Markov chain with a unique stationary distribution π, and when $X_1 \sim \pi$, $\{(X_t, Y_t), t \geq 1\}$ is ergodic. As for inference, we can first estimate $\theta = (\delta, \alpha, \beta)^T$ for r fixed and calculate the profile likelihood as a function of r. Then r is estimated by choosing the one that maximizes the profile likelihood, and $\hat{\theta}$ can be obtained correspondingly. Moreover, if we assume r is known and $(\alpha+\beta)^2 + \beta^2/r < 1$, then under (L0), the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent and asymptotically normal. The proof can be found in Appendix B.4.

4.2. Nonlinear dynamic models

It is possible to generalize (4.1) to nonlinear dynamic models. While the theory developed in Sections 2 and 3 can be applied to many general nonlinear dynamics, we will elaborate on one approach that is based on the idea of spline basis functions, see for example, Ruppert, Wand, and Carroll (2003). In this framework, the model specification is

$$Y_t|\mathcal{F}_{t-1} \sim p(y|\eta_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1} + \sum_{k=1}^{K} \beta_k (Y_{t-1} - \xi_k)^+, \tag{4.8}$$

where $K \in \mathbb{N}_0$, $\delta > 0, \alpha, \beta \geq 0, \beta_1, \ldots, \beta_K$ are parameters, $\{\xi_k\}_{k=1}^K$ are the so-called *knots*, and x^+ is the positive part of x. In particular, when $K = 0$, (4.8) reduces to the linear model (4.1). It is easy to see that (4.8) is a special case of (2.2) by taking $g_\theta(x, y) = \delta + \alpha x + \beta y + \sum_{k=1}^K \beta_k (y - \xi_k)^+$, where $\theta = (\delta, \alpha, \beta, \beta_1, \ldots, \beta_K)^T$.

Being piecewise linear in Y_{t-1}, (4.8), in principle, provides a wide and flexible family of nonlinear dynamics. In each of the pieces segmented by the knots, (4.8) has INGARCH-like dynamics. For example, if $Y_{t-1} \in [\xi_s, \xi_{s+1})$ for some $s < K$, then $X_t = (\delta - \sum_{k=1}^s \beta_k \xi_k) + \alpha X_{t-1} + (\beta + \sum_{k=1}^s \beta_k) Y_{t-1}$. This can be viewed as one of the generalizations (e.g., Samia and Chan (2010)) to the threshold autoregressive model (Tong (1990)). Wang et al. (2014) proposed the self-excited threshold Poisson autoregression model as a generalization to the Poisson INGARCH (4.5) is an attempt to model negative autocorrelation. Model (4.8) turns out to be a special case of it, in that the same α is shared across regimes. According to Propositions 1 and 2, we can establish the stability properties of (4.8).

Proposition 5. *If* (4.8) *holds with* $\alpha + \beta < 1, \beta + \sum_{k=1}^s \beta_k \geq 0$ *and* $\alpha + \beta + \sum_{k=1}^s \beta_k < 1$ *for* $s = 1, \ldots, K$, *then* $\{X_t\}$ *is geometric moment contracting and has a unique stationary distribution* π. *Moreover if* $X_1 \sim \pi$, *then* $\{(X_t, Y_t), t \geq 1\}$ *is ergodic.*

We consider inference for this model. If we assume the knots $\{\xi_k\}_{k=1}^K$ are known for K fixed, then the parameter vector $\theta = (\delta, \alpha, \beta, \beta_1, \ldots, \beta_K)^T$ can be estimated by maximizing the conditional log-likelihood function, which is available according to (3.1). The number of knots K can be selected by virtue of an information criteria, such as AIC and BIC. As for the locations of knots, there are different strategies one can adopt for choosing them. One method is to place the knots at the $\{j/(K+1), j = 1, \ldots, K\}$ quantiles of the population, which can be estimated from the data. A second method is to choose the locations that

maximize the log likelihood. We will apply both procedures to datasets in the next section.

To study the asymptotic behavior of the estimates, first note that by iterating the recursion,

$$X_t = \frac{\delta}{1-\alpha} + \beta \sum_{i=0}^{\infty} \alpha^i Y_{t-1-i} + \sum_{k=1}^{K} \beta_k \sum_{i=0}^{\infty} \alpha^i (Y_{t-1-i} - \xi_k)^+$$
$$= \frac{\delta}{1-\alpha} + \sum_{i=0}^{\infty} \alpha^i \{\beta Y_{t-1-i} + \sum_{k=1}^{K} \beta_k (Y_{t-1-i} - \xi_k)^+\}. \tag{4.9}$$

This defines the function g_∞^θ as in $X_t = g_\infty^\theta(Y_{t-1}, Y_{t-2}, \ldots)$ and also verifies assumptions (A1)−(A3). Hence, to apply Theorem 3, we need only impose some regularity assumptions on (4.8).

(NL1) θ_0 is an interior point in the parameter space Θ, which is a compact subset of the parameter set satisfying the conditions in Proposition 5.

(NL1) $\mathrm{E}[Y_1 \sup_{\theta\in\Theta} B^{-1}((\delta/(1-\alpha)) + \sum_{i=0}^{\infty} \alpha^i \{\beta Y_{t-1-i} + \sum_{k=1}^{K} \beta_k (Y_{t-1-i} - \xi_k)^+\})] < \infty.$

(NL2) $\mathrm{E}[B'(\eta_1(\theta_0))\{\partial \eta_1(\theta)/\partial \theta_i)\}^2|_{\theta=\theta_0}] < \infty$, for $i = 1, \ldots, K+3$.

Sufficient conditions for assumptions (NL1) and (NL2) can be established similarly to those given in Remarks 1 and 2 in Appendix B.

Theorem 4. *If* (4.8) *holds with known placement of the knots, and* (NL0)−(NL2) *hold, then the maximum likelihood estimator* $\hat{\theta}_n$ *is strongly consistent and*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega^{-1}), \quad as \quad n \to \infty,$$

where $\Omega = E\{B'(\eta_t)\dot{\eta}_t \dot{\eta}_t^T\}$.

We use the Poisson nonlinear dynamic model as an example of the results; see Section 5 for implementation of the estimation procedure. The model is

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{Pois}(\lambda_t), \quad \lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1} + \sum_{k=1}^{K} \beta_k (Y_{t-1} - \xi_k)^+. \tag{4.10}$$

Under the conditions of Proposition 5 and Theorem 4, $\{(\lambda_t, Y_t), t \geq 1\}$ is a stationary and ergodic process, and the estimates are strongly consistent and asymptotically normal. In practice the covariance matrix of the estimates can be obtained by recursively applying

$$\frac{\partial \lambda_t}{\partial \theta} = \left(1, \lambda_{t-1}, Y_{t-1}, (Y_{t-1} - \xi_1)^+, \ldots, (Y_{t-1} - \xi_K)^+\right)^T + \alpha \frac{\partial \lambda_{t-1}}{\partial \theta}.$$

Another example is the Poisson exponential autoregressive model proposed by Fokianos, Rahbek, and Tjøstheim (2009). Here

$$Y_t|\mathcal{F}_{t-1} \sim \text{Pois}(\lambda_t), \quad \lambda_t = (\alpha_0 + \alpha_1 \exp\{-\gamma\lambda_{t-1}^2\})\lambda_{t-1} + \beta Y_{t-1}, \tag{4.11}$$

where $\alpha_0, \alpha_1, \beta, \gamma > 0$ are parameters. If $\alpha_0 + \alpha_1 + \beta < 1$, then model (4.11) belongs to the class of models at (2.2) and hence enjoys the stability properties stated in Propositions 1 and 2. As for the inference of the model, see Fokianos, Rahbek, and Tjøstheim (2009).

5. Numerical Results

The performance of the estimation procedure for the Poisson nonlinear dynamic model is illustrated in a simulation study. The MLE was obtained by optimizing the log-likelihood function (3.1) using a Newton-Raphson method. Simulation results of the Poisson INGARCH can be found in Fokianos, Rahbek, and Tjøstheim (2009). Other models, including the negative binomial linear and nonlinear dynamic models and the exponential autoregressive model (4.11) were applied to two datasets, and tools for checking goodness of fit were considered.

5.1. Simulation for the nonlinear model

A 1-knot nonlinear dynamic model was simulated according to

$$Y_t|\mathcal{F}_{t-1} \sim \text{Pois}(\lambda_t), \quad \lambda_t = 0.5 + 0.5\lambda_{t-1} + 0.4Y_{t-1} - 0.2(Y_{t-1} - 5)^+$$

with different sample sizes. Each sample size and parameter configuration was replicated 1,000 times. For each realization, the first 500 simulated observations were discarded as burn-in in order to let the process reach its stationary regime. We first estimated the parameters assuming the true underlying model (4.8) with only one knot at 5. The means and standard errors of the estimates from all 1,000 runs are summarized in Table 1, and the histograms of the estimates are depicted in Figure 1. The performance of these estimates is reasonably good and consistent with the theory in Theorem 4. As for estimating the parameters without knowing the location of the knots, the corresponding results of the MLE obtained by fitting a 1-knot model to all the 1,000 replications are summarized in Table 2. Here the locations of the knots were determined by sample quantiles. The performance of the maximum likelihood estimates of β and β_1 is not as good as in the known knot case, but the overall model performance, as reflected in the computation of the scoring rules (described in the next section), is competitive with the known knot case.

For the problem of selecting the number of knots using an information criterion, simulations with different sample sizes were implemented; model selection

Table 1. Estimation results for 1-knot model with known knot location.

	δ	α	β	β_1	n
True	0.5	0.5	0.4	-0.2	
Estimates	0.5596	0.4861	0.3990	-0.2009	500
s.e.	(0.0087)	(0.0030)	(0.0026)	(0.0051)	
Estimates	0.5265	0.4944	0.3991	-0.2016	1,000
s.e.	(0.0041)	(0.0016)	(0.0013)	(0.0025)	

Table 2. Estimation for 1-knot model with unknown knot location.

	δ	α	β	β_1	n
True	0.5	0.5	0.4	-0.2	
Estimates	0.5387	0.4852	0.4187	-0.1614	500
s.e.	(0.0089)	(0.0030)	(0.0031)	(0.0047)	
Estimates	0.5002	0.4943	0.4197	-0.1679	1,000
s.e.	(0.0042)	(0.0016)	(0.0015)	(0.0023)	

Table 3. Model selection of 1-knot simulation.

Criteria	0 knot	1 knot	2 knots	3 knots	≥ 4 knots	n
AIC	34.3%	37.6%	20.9%	5.2%	2.0%	500
BIC	80.5%	18.8%	0.6%	0.1%	0	
AIC	12.4%	45.0%	29.9%	8.3%	4.4%	1,000
BIC	59.4%	38.4%	2.0%	0.2%	0	

results are summarized in Table 3. Numbers in the table stand for the proportion of times that each particular model was selected in the 1,000 runs. For AIC, the 1-knot model was selected most often followed by a 2-knot model, at least in the cases when n =1,000.

In interpolating the nonlinear dynamic of λ_t by a piecewise linear function, we plot in Figure 2 the fitted functions $\hat{\beta}y + \sum_{k=1}^{K} \hat{\beta}_k (y - \hat{\xi}_k)^+$ for each run of the simulations against its true form $0.4y - 0.2(y-5)^+$. In the graph, the piecewise linear function fitted by the 1-knot model is closest to the true curve.

5.2. Two data applications

1. Number of transactions of Ericsson stock

Both linear and nonlinear dynamic models were employed to fit the number of transactions per minute for the stock Ericsson B during July 2nd, 2002 which consists of 460 observations. Figure 3 plots the data and the autocorrelation function. The positive dependence displayed in the data suggests the application of the models in our study.

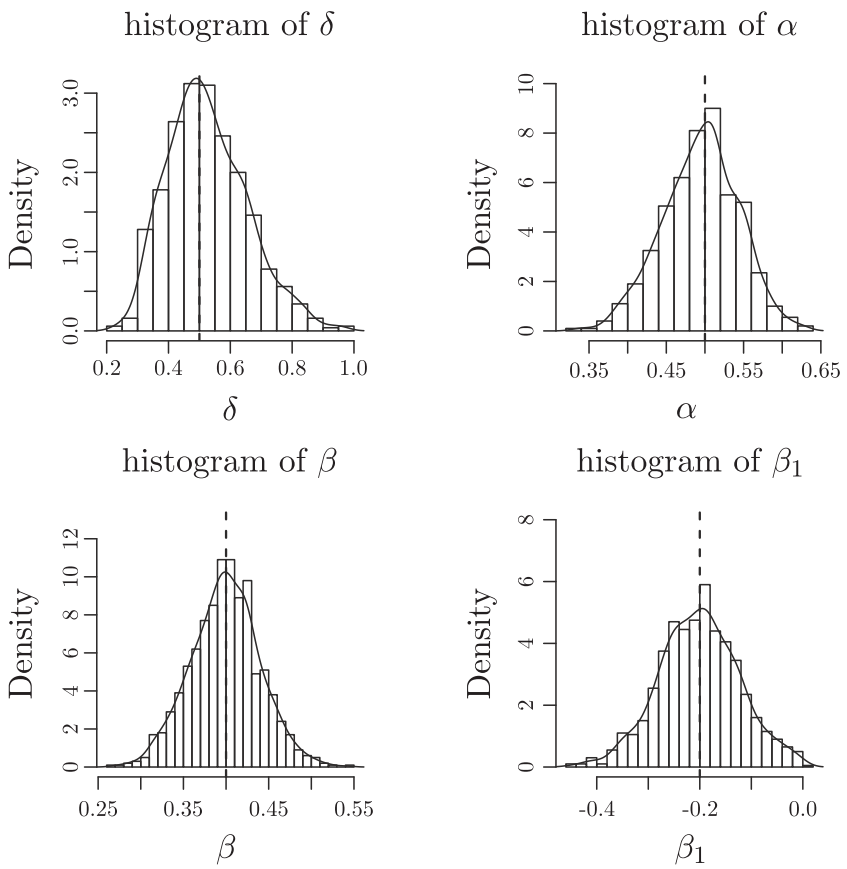

Figure 1. Histograms of the 1-knot model with sample size 1,000 assuming the knot is known. The overlaying curves are the density estimates and the dashed vertical lines represent the true values of the parameters.

By computing the MLE of the parameters, the fitted Poisson INGARCH model is given by

$$\hat{\lambda}_t = \underset{(0.1000)}{0.2912} + \underset{(0.0242)}{0.8312}\hat{\lambda}_{t-1} + \underset{(0.0188)}{0.1395}Y_{t-1},$$

and the fitted NB-INGARCH model is

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{NB}(8, \hat{p}_t), \quad \hat{X}_t = \underset{(0.1406)}{0.2676} + \underset{(0.0350)}{0.8447}\hat{X}_{t-1} + \underset{(0.0274)}{0.1282}Y_{t-1},$$

where $\hat{X}_t = 8(1-\hat{p}_t)/\hat{p}_t$. The standard deviations in the parentheses were calculated according to the remark after Theorem 2.

As for the Poisson nonlinear dynamic model, AIC and BIC were used to help select the number of knots among 0 to 5; the values are reported in Table 4. The

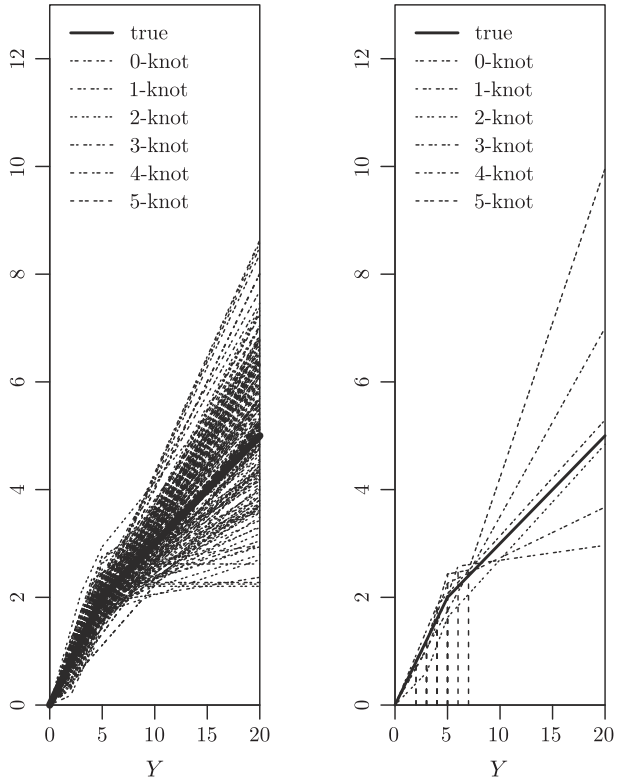

Figure 2. Left: the thick black curve is the true function $0.4y - 0.2(y - 5)^+$, and the other curves are the piecewise linear functions fitted in each simulation where the number of knots K is selected via AIC; Right: for each value of K, we plot the fitted curve from one specific run that chooses the particular number of knots.

Table 4. Model selection results for Ericsson data.

	0-knot	1-knot	2-knot	3-knot	4-knot	5-knot
LogL	-1433.19	-1431.21	-1431.08	-1430.58	**-1429.65**	-1431.12
AIC	2874.38	**2872.41**	2874.17	2875.17	2875.30	2880.25
BIC	**2890.90**	2893.07	2898.95	2904.08	2908.35	2917.43

fitted 1-knot Poisson model, which had the smallest AIC, is given by

$$\hat{\lambda}_t = \underset{(0.1884)}{0.5837} + \underset{(0.0241)}{0.8319}\hat{\lambda}_{t-1} + \underset{(0.0295)}{0.0906}Y_{t-1} + \underset{(0.0373)}{0.0722}(Y_{t-1} - 9)^+.$$

The AIC values of the 2-knot and 3-knot models are both close to that of the 1-knot model, and therefore are used as a basis for comparison with the minimum AIC model. These models are given by $\hat{\lambda}_t = 0.5519 + 0.8326\hat{\lambda}_{t-1} + 0.0961Y_{t-1} +$

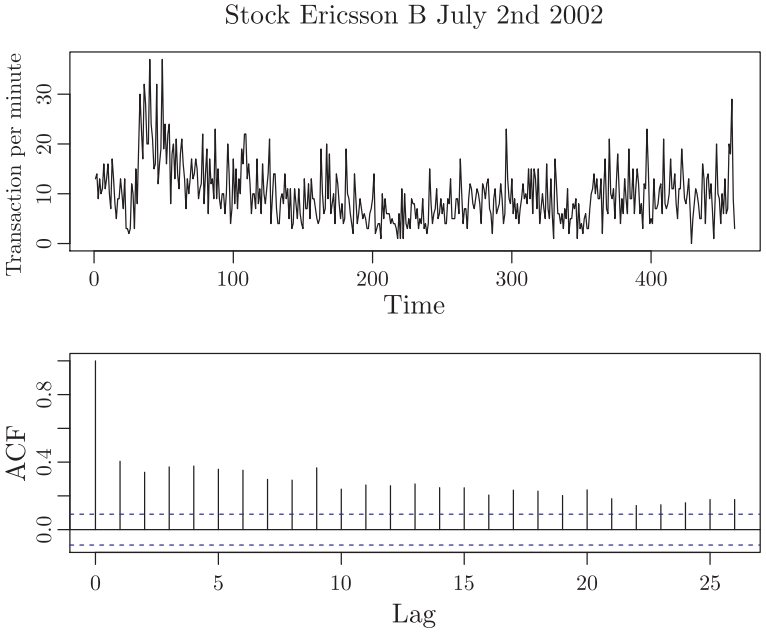

Figure 3. Top: Number of transactions per minute of the stock Ericsson B during July 2nd 2002; Bottom: ACF of the data.

$0.0154(Y_{t-1}-7)^+ + 0.0559(Y_{t-1}-11)^+$ and $\hat{\lambda}_t = 0.3614 + 0.8361\hat{\lambda}_{t-1} + 0.1206Y_{t-1} + 0.0433(Y_{t-1}-6)^+ - 0.0914(Y_{t-1}-9)^+ + 0.0914(Y_{t-1}-13)^+$, respectively.

From our model checking, the negative binomial INGARCH model seems to outperform the Poisson-based models. This could be explained by the over-dispersion exhibited by the data, since the mean and variance are 9.91 and 32.84, respectively. To this end, we fit the nonlinear negative binomial models and selected the number of knots by minimizing the AIC. The AIC value of a 1-knot model was the second smallest among all the candidates, with 2,674.69 compared to the smallest value 2,674.04, which is attained by the negative binomial fitted INGARCH. The fitted 1-knot negative binomial nonlinear model is given by $Y_t|\mathcal{F}_{t-1} \sim \mathrm{NB}(8, \hat{p}_t)$, where $\hat{X}_t = 8(1-\hat{p}_t)/\hat{p}_t$ follows

$$\hat{X}_t = \underset{(0.2559)}{0.4931} + \underset{(0.0350)}{0.8444}\hat{X}_{t-1} + \underset{(0.0412)}{0.0903}Y_{t-1} + \underset{(0.0546)}{0.0603}(Y_{t-1}-9)^+.$$

Here the locations of knots for the nonlinear dynamic model were estimated by the corresponding sample quantiles. We also tried estimating the knots by maximizing the likelihood and, in this application, the results by both methods were nearly identical. The exponential autoregressive model (4.11) is also applied to this dataset by Fokianos, Rahbek, and Tjøstheim (2009); it is given by

$$\hat{\lambda}_t = (0.8303 + 7.030\exp\{-0.1675\hat{\lambda}_{t-1}^2\})\hat{\lambda}_{t-1} + 0.1551Y_{t-1}.$$

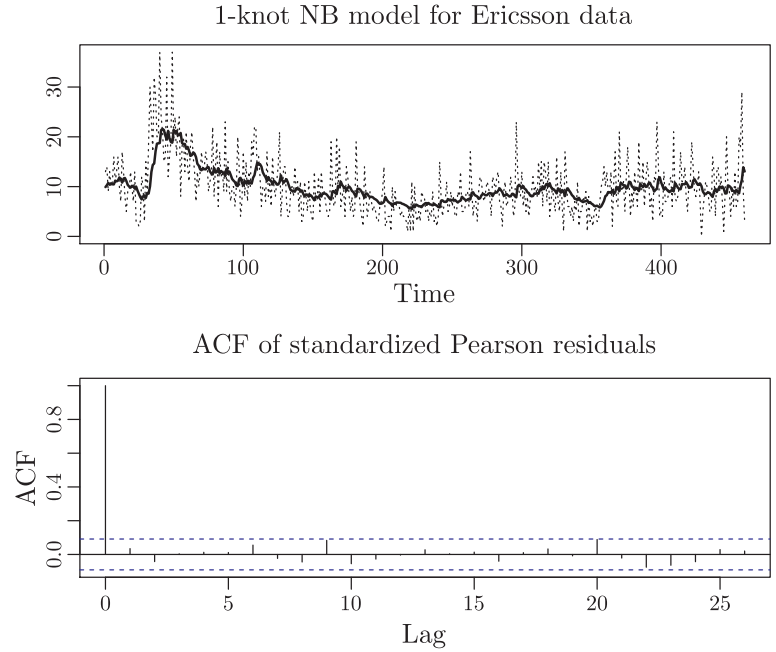

Figure 4. Top: Dotted curve represents the number of transactions of Ericsson stock, and the black curve is the fitted conditional mean process by 1-knot NB-based model; Bottom: ACF of the standardized Pearson residuals.

(0.0232) (3.0732) (0.0592) (0.0218)

To assess the adequacy of fit, we consider an array of graphical and quantitative diagnostic tools for time series, some of which are specifically designed for time series of counts. Readers can refer to Davis, Dunsmuir, and Streett (2003) and Jung and Tremayne (2011) for a comprehensive treatment of the tools. We first consider the standardized Pearson residuals $e_t = (Y_t - \mathrm{E}(Y_t|\mathcal{F}_{t-1}))/\sqrt{\mathrm{Var}(Y_t|\mathcal{F}_{t-1})}$ which can be obtained by replacing the population quantities by their estimated counterparts. If the model is correctly specified, then the residuals $\{\hat{e}_t\}$ should be a white noise sequence with constant variance. All the models considered above give very similar fitted conditional mean processes and the standardized Pearson residuals appear to be white. Figure 4 displays the fitted result for the 1-knot negative binomial model.

Another model check uses probability integral transform (PIT); when the underlying distribution is continuous, the PIT is standard uniform. If the underlying distribution is discrete, some adjustments are required and the so-called randomized PIT is introduced by perturbing the step function characteristic of the CDF of discrete random variables (see Brockwell (2007)). More recently,

Table 5. Quantitative model checking for Ericsson data.

Model	log likelihood	p-value of PIT	LS	QS	RPS
Poisson INGARCH	-1433.19	$< 10^{-5}$	3.1167	-0.0576	2.6883
NB INGARCH	-1332.02	0.7386	2.8958	-0.0671	2.6063
1-knot Poisson model	-1431.21	$< 10^{-5}$	3.1123	-0.0573	2.6848
2-knot Poisson model	-1431.08	$< 10^{-5}$	3.1121	-0.0575	2.6843
3-knot Poisson model	-1430.58	$< 10^{-5}$	3.1110	-0.0580	2.6779
1-knot NB model	**-1331.34**	0.8494	**2.8942**	**-0.0671**	**2.6021**
Exp-auto model	-1448.69	$< 10^{-5}$	3.1504	-0.0600	2.6924

Czado, Gneiting, and Held (2009) proposed a non-randomized version of PIT as an alternative adjustment. Since it usually gives the same conclusion for model checking, we do not provide the non-randomized version here. For any t, the randomized PIT is

$$\tilde{u}_t := F_t(Y_t - 1) + \nu_t\big[F_t(Y_t) - F_t(Y_t - 1)\big],$$

where $\{\nu_t\}$ is a sequence of iid uniform $(0, 1)$ random variables, and $F_t(\cdot)$ is the predictive cumulative distribution. Here $F_t(\cdot)$ is simply the CDF of a Poisson or a negative binomial distribution. If the model is correct, then $\tilde{u}_t$ is an iid sequence of uniform $(0, 1)$ random variables. Jung and Tremayne (2011) reviewed several ways to depict this and we adopt their method in our study. To test if the PIT is uniform $(0, 1)$, the histograms of PIT from different models were plotted and a Kolmogorov-Smirnov test was carried out. The results are summarized in Figure 5, and the p-values are reported in Table 5. The two negative binomial-based models pass the PIT test, while none of the Poisson-based models does. This observation could again be explained by the over-dispersion phenomenon of the data.

To measure the power of predictions by models, various scoring rules have been proposed in literature, see e.g., Czado, Gneiting, and Held (2009) and Jung and Tremayne (2011). Most of them are computed as the average of quantities related to predictions and take the form $(n-1)^{-1}\sum_{t=2}^{n} s(F_t(Y_t))$ where $F_t(\cdot)$ is the CDF of the prediction distribution and $s(\cdot)$ denotes some scoring rule. We calculated three scoring rules for the fitted models: logarithmic score (LS), quadratic score (QS) and ranked probability score (RPS). For definitions, see Jung and Tremayne (2011). Table 5 summarizes these scores for all of the fitted models. As seen from the table, most of the diagnostic tools favor the one-knot negative binomial model for the Ericsson data.

2. Return times of extreme events of Goldman Sachs Group (GS) stock

We constructed a time series based on daily log-returns of Goldman Sachs Group (GS) stock from May 4th, 1999 to March 16th, 2012. We first calculated

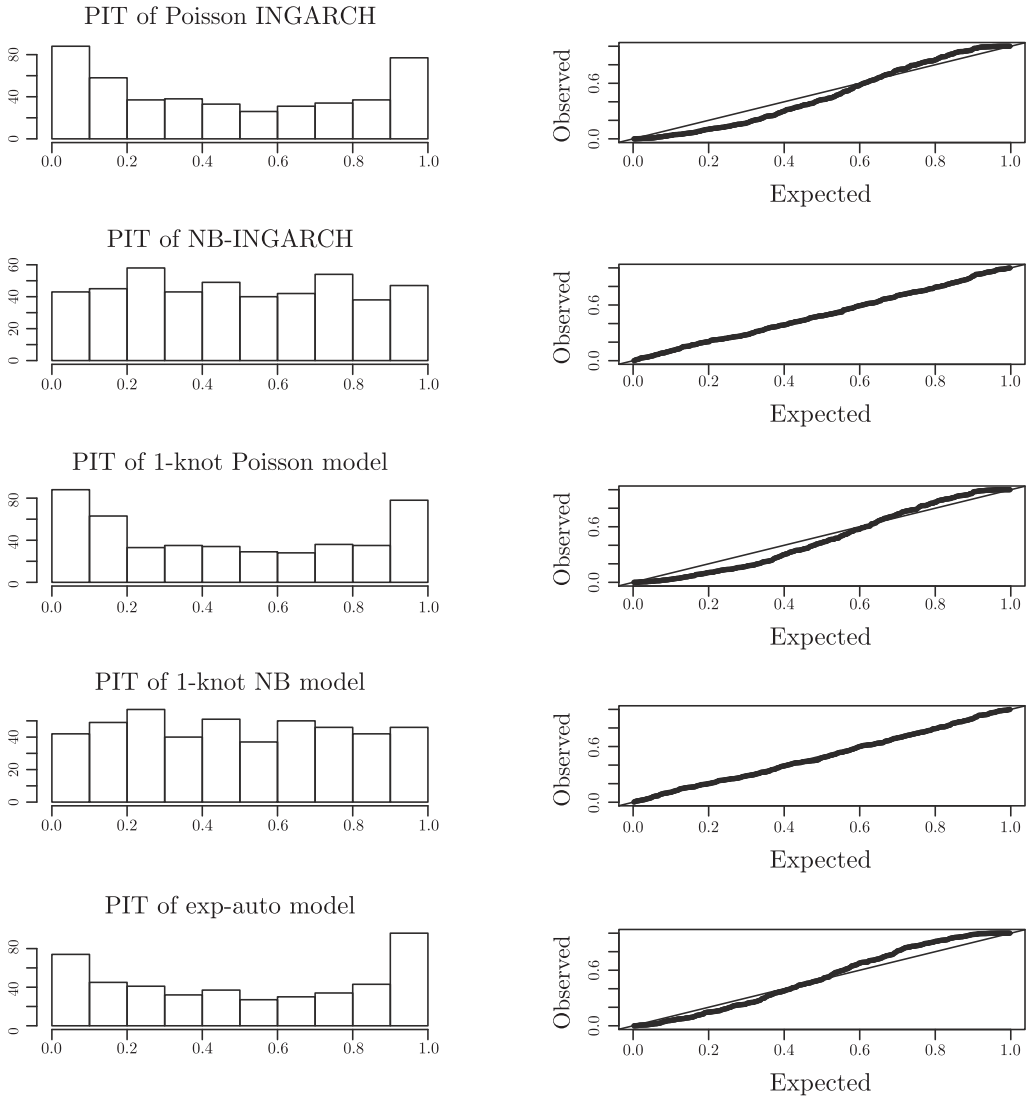

Figure 5. Left: histograms of randomized PIT's for all of the models fitted to the Ericsson stock data; Right: QQ-plots of $\tilde{u}_t$ against standard uniform distribution for the corresponding models, where the straight line is the 45° line with zero intercept.

the hitting times, $\tau_1, \tau_2, \ldots$, for which the log-returns of GS stock falls outside the 0.05 and 0.95 quantiles of the data. The discrete time series of interest is the return (or inter-arrival) times $Y_t = \tau_t - \tau_{t-1}$. If the data are in fact iid, or do not exhibit clustering of large values, then the Y_t's should be independent and geometrically distributed with probability of success $p = 0.1$ (Chang (2010)). Figure 6 plots the return times of the stock, and the ACF and histogram of the

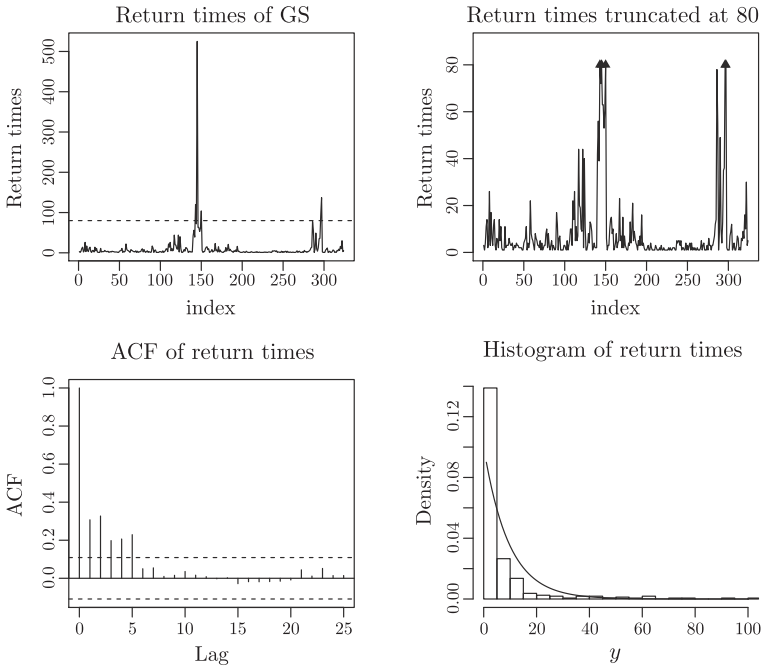

Figure 6. Top left: Return times of GS stock, the dashed horizontal line locates at 80; Top right: Return times truncated at 80 in order to ameliorate the visual effect of the five large observations that are represented by solid triangles; Bottom left: ACF of the return times; Bottom right: Histogram of the return times, where the curve overlaid is the density function of a geometric distribution with $p = 0.1$.

return times. In order to ameliorate the visual effect of some extremely large observations, the time series is also plotted in the top right panel of Figure 6 on a reduced vertical scale in which it is truncated at 80, and the five observations that are affected are depicted by solid triangles.

To explore this time series, the geometric INGARCH (negative binomial INGARCH (4.7) with $r = 1$), and the 1-knot and 2-knot geometric-based models were fitted to the data. The number of knots for the nonlinear dynamic models was chosen by minimizing the AIC, and the locations of knots were estimated by maximizing the likelihood based on a grid search. A constraint was imposed: there should be at least 30 observations in each of the regimes segmented by the knots in order to guarantee that there are sufficient observations to obtain quality estimates of the parameters. The sample quantile method for estimating knot locations did not perform as well.

Since $Y_t \geq 1$ for any t, we used a version of the geometric distribution that counts the total number of trials, instead of the failures. In particular, the fitted

Table 6. Quantitative model checking for GS return times.

Model	log likelihood	p-value of PIT	LS	QS	RPS
Poisson INGARCH	-2681.06	$< 10^{-5}$	8.2842	-0.0675	4.1373
Geom INGARCH	-857.73	0.2581	2.6477	-0.1436	3.4100
3-knot Poisson model	-2670.33	$< 10^{-5}$	8.2510	-0.0693	4.1400
1-knot Geom model	-857.58	0.3988	2.6472	**-0.1436**	3.4041
2-knot Geom model	**-857.42**	0.2006	**2.6468**	-0.1435	**3.3939**

1-knot geometric-based model is given by $Y_t - 1|\mathcal{F}_{t-1} \sim \text{Geom}(p_t)$, where

$$X_t = 0.5042 + 0.4729X_{t-1} + 0.5271(Y_{t-1} - 1) - 0.0526(Y_{t-1} - 5)^+,$$

and the fitted 2-knot geometric-based model is

$$X_t = 0.5414 + 0.4531X_{t-1} + 0.5469Y_{t-1} - 0.2333(Y_{t-1} - 9)^+ + 0.2332(Y_{t-1} - 18)^+,$$

where $X_t = (1 - p_t)/p_t$. In both models, $\hat{\alpha} + \hat{\beta}$ is very close to unity, so the estimated parameters are close to the boundary of the parameter space. This is similar to the integrated GARCH (IGARCH) model in which $\alpha + \beta = 1$. In our application, the mean of the time series of return times is about 10, while the variance is 1,101. A simulation according to the fitted model gave mean and median close to those of the data, but the variance of the simulated data was extraordinarily large, resembling the feature of the observed data. Because, the fitted models are still stationary, the parameters no longer satisfy the conditions specified in Theorem 4 that ensure a finite variance.

The fitted geometric-based models are capable of capturing the high volatility part of the data. Their standardized Pearson residuals were also calculated and appear to be white. Results of the PIT test are depicted in Figure 7, and the prediction scores and the p-values of the PIT test are summarized in Table 6. Two Poisson-based models are also included for comparison and, as expected, they do not perform as well as the geometric-based models.

Acknowledgement

This research is supported in part by NSF grant DMS-1107031.

Appendix A. Properties of the exponential family

An important property of the one-parameter exponential family is stochastic monotonicity. A random variable X is said to be stochastically smaller than a random variable Y (written as $X \leq_{ST} Y$) if $F(x) \geq G(x)$ for all x, where $F(x)$ and $G(x)$ are the cumulative distribution functions of X and Y respectively. We refer readers to Yu (2009) for the related theory.

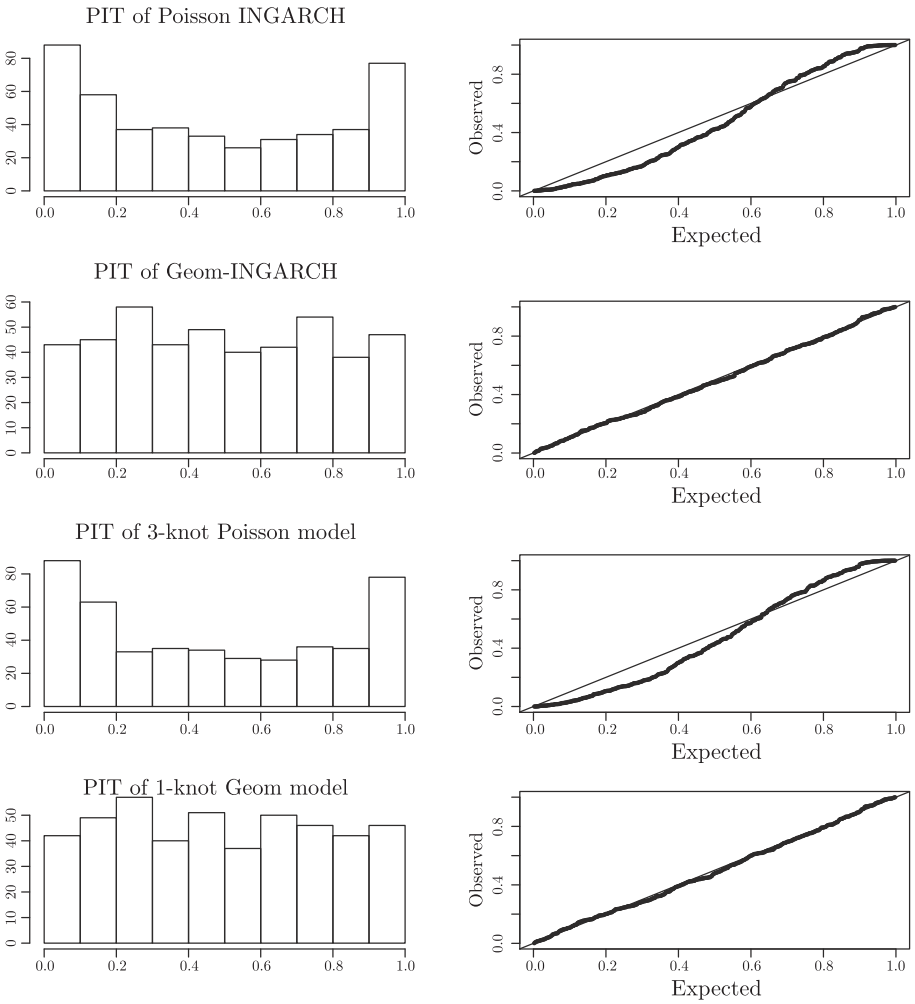

Figure 7. Left: histograms of randomized PIT's for the models fitted to GS return times; Right: QQ-plots of $\tilde{u}_t$ against standard uniform distribution for the corresponding models, where the straight line is the 45° line with zero intercept.

Proposition A.1. Suppose random variables Y' and Y'' follow distributions belonging to the one-parameter exponential family (2.1) with the same A, h, and μ, but with natural parameters η' and η'', respectively. If $\eta' \leq \eta''$, then Y' is stochastically smaller than Y''.

Proof. Let the probability density functions of Y' and Y'' be $p(y|\eta')$ and $p(y|\eta'')$, respectively. Then their log ratio is

$$l(y) = \log \frac{p(y|\eta')}{p(y|\eta'')} = \log \frac{\exp\{\eta' y - A(\eta')\}h(y)}{\exp\{\eta'' y - A(\eta'')\}h(y)}$$

$$= y(\eta' - \eta'') + [A(\eta'') - A(\eta')],$$

a concave function in y. From Definition 2 in Yu (2009), Y' is log concave relative to Y'', $Y' \leq_{lc} Y''$. As $A(\eta)$ is increasing in η, $\lim_{y\downarrow 0} l(y) = A(\eta'') - A(\eta') \geq 0$ for continuous $p(y|\eta)$, and $p(0|\eta')/p(0|\eta'') \geq 1$ for discrete $p(y|\eta)$. Hence, according to Theorem 1 in Yu (2009), Y' is stochastically smaller than Y'', $Y' \leq_{ST} Y''$.

Let F_x be the cumulative distribution function of $p(y|\eta)$ in (2.1) with $x = B(\eta)$, and its inverse be $F_x^{-1}(u) := \inf\{t \geq 0 : F_x(t) \geq u\}$ for $u \in [0, 1]$.

Proposition A.2. Suppose U is uniform $(0, 1)$, and take $Y' = F_{x'}^{-1}(U)$ and $Y'' = F_{x''}^{-1}(U)$, where $x' = B(\eta')$ and $x'' = B(\eta'')$. Then $\mathrm{E}|Y' - Y''| = |x' - x''|$.

Proof. It follows from the construction of Y' and Y'' that they follow the one-parameter exponential family (2.1) with natural parameters η' and η'' respectively, and $\mathrm{E}Y' = x'$, $\mathrm{E}Y'' = x''$. If $x' \leq x''$, then Y' is stochastically smaller than Y'' by virtue of Proposition A.1. It follows that $F_{x'}^{-1}(\theta) \leq F_{x''}^{-1}(\theta)$ for $\theta \in (0, 1)$, i.e., $Y' \leq Y''$. This implies $\mathrm{E}|Y' - Y''| = \mathrm{E}(Y'' - Y') = x'' - x'$. Similarly if $x' \geq x''$, then $\mathrm{E}|Y' - Y''| = x' - x''$. Hence we have $\mathrm{E}|Y' - Y''| = |x' - x''|$.

Appendix B. Linear dynamic models and examples

At (4.1), note that by recursion we have, for all t,

$$X_t(\theta) = \frac{\delta}{1-\alpha} + \beta \sum_{k=0}^{\infty} \alpha^k Y_{t-1-k}. \tag{B.1}$$

It follows that $X_t(\theta) \geq x^* = \delta/(1-\alpha)$ since Y_t only takes non-negative values.

B.1. Hessian matrix

In addition to the score function derived in Section 4.1, the Hessian matrix can be found by taking derivatives of the score function:

$$H_n(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T} = \sum_{t=1}^{n} [-B'(\eta_t(\theta)) \frac{\partial \eta_t(\theta)}{\partial\theta} \frac{\partial \eta_t(\theta)}{\partial\theta^T} + \{Y_t - B(\eta_t(\theta))\} \frac{\partial^2 \eta_t(\theta)}{\partial\theta\partial\theta^T}],$$

where

$$\begin{aligned}\frac{\partial^2 \eta_t}{\partial\theta\partial\theta^T} = \Bigg(& \frac{B''(\eta_t)}{(B'(\eta_t))^2} \frac{\partial \eta_t}{\partial\theta} \quad \frac{B'(\eta_{t-1})B'(\eta_t)}{(B'(\eta_t))^2} \frac{\partial \eta_{t-1}}{\partial\theta} - \frac{B'(\eta_{t-1})B''(\eta_t)}{(B'(\eta_t))^2} \frac{\partial \eta_t}{\partial\theta} \\ & \frac{-Y_{t-1}B''(\eta_t)}{(B'(\eta_t))^2} \frac{\partial \eta_t}{\partial\theta} \Bigg) + (0 \quad 1 \quad 0)^T \frac{B'(\eta_{t-1})}{B'(\eta_t)} \frac{\partial \eta_{t-1}}{\partial\theta^T} + \alpha \frac{B''(\eta_{t-1})B'(\eta_t)}{(B'(\eta_t))^2} \\ & \frac{\partial \eta_{t-1}}{\partial\theta} \frac{\partial \eta_{t-1}}{\partial\theta^T} - \alpha \frac{B'(\eta_{t-1})B''(\eta_t)}{(B'(\eta_t))^2} \frac{\partial \eta_t}{\partial\theta} \frac{\partial \eta_t}{\partial\theta^T} + \alpha \frac{B'(\eta_{t-1})}{B'(\eta_t)} \frac{\partial^2 \eta_{t-1}}{\partial\theta\partial\theta^T}.\end{aligned}$$

B.2. Proofs of Remarks 1 and 2

Proof. Remark 1 can ben seen by noting that $X_1(\theta) \le \delta_U/\epsilon + \sum_{k=1}^{\infty}(1-\epsilon)^k Y_{1-k}$. For Remark 2, the most difficult case is the derivative with respect to $\theta_2 = \alpha$ and we only give its proof, since the arguments for δ and β are similar. First note that

$$\mathrm{E}\{B'(\eta_1(\theta_0))\big(\frac{\partial \eta_1(\theta_0)}{\partial \alpha}\big)^2\} = \mathrm{E}\{\frac{1}{B'(\eta_1)}\big(\frac{\partial B(\eta_1)}{\partial \alpha}\big)^2\} \le \frac{1}{\underline{c}}\mathrm{E}\{\frac{\partial B(\eta_1)}{\partial \alpha}\}^2,$$

where $\partial B(\eta_1)/\partial\alpha = \delta/(1-\alpha)^2 + \beta\sum_{k=1}^{\infty} k\alpha^{k-1}Y_{-k}$. On account of stationarity, one can show that

$$\begin{aligned}\mathrm{E}\big(\sum_{k=1}^{\infty} k\alpha^{k-1}Y_{-k}\big)^2 \le \{\gamma_Y(0) + \frac{2\gamma_Y(1)}{1-\alpha(\alpha+\beta)}\}\sum_{k=1}^{\infty} k^2\alpha^{2k-2} \\ + \frac{2\alpha\gamma_Y(1)}{1-\alpha^2(\alpha+\beta)^2}\sum_{k=1}^{\infty} k\alpha^{2k-2} + \mu^2\big(\sum_{k=1}^{\infty} k\alpha^{k-1}\big)^2 < \infty,\end{aligned}$$

where $\mu = \mathrm{E}Y_t < \infty$. Hence $\mathrm{E}[B'(\eta_1(\theta_0))\{\partial\eta_1(\theta_0)/\partial\alpha\}^2] < \infty$ if $\gamma_Y(0) < \infty$.

B.3. More on the Poisson INGARCH

The Poisson INGARCH(1, 1) model is

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{Pois}(\lambda_t), \quad \lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1},$$

where $\delta > 0, \alpha, \beta \ge 0$ are parameters. When $\alpha + \beta < 1$, $\{(Y_t, \lambda_t), t \ge 1\}$ is an ergodic stationary process and the MLE $\hat{\theta}_n$ is strongly consistent and asymptotically normal by Theorem 3. To see this, we need only verify assumptions (L1) and (L2). From Fokianos, Rahbek, and Tjøstheim (2009), we have $\gamma_Y(0) = \{1-(\alpha+\beta)^2+\beta^2\}/\{1-(\alpha+\beta)^2\}$ and $\gamma_Y(h) = \mu C(\theta)(\alpha+\beta)^{h-1}$ for $h \ge 1$, where $\mu = \mathrm{E}Y_t = \delta/(1-\alpha-\beta)$ and $C(\theta)$ is a positive constant dependent on θ. Hence by the Monotone Convergence Theorem, we have

$$\begin{aligned}\mathrm{E}[Y_1 \log\{\delta_U/\epsilon + \sum_{k=1}^{\infty}(1-\epsilon)^k Y_{1-k}\}] &\le \mathrm{E}[Y_1\{\delta_U/\epsilon + \sum_{k=1}^{\infty}(1-\epsilon)^k Y_{1-k}\}] \\ &= \frac{\delta_U}{\epsilon}\mathrm{E}Y_1 + \sum_{k=1}^{\infty}(1-\epsilon)^k \mathrm{E}Y_1Y_{1-k} \\ &= \mu\frac{\delta_U}{\epsilon} + \sum_{k=1}^{\infty}(1-\epsilon)^k\{\gamma_Y(k) + \mu^2\} < \infty.\end{aligned}$$

Hence assumption (L1) holds according to Remark 1. Now $B(\eta_t) = \lambda_t \ge \lambda^* := \delta/(1-\alpha)$ for all t, so $A''(\eta_t) = e^{\eta_t}$ is bounded away from 0, so (L2) holds according to Remark 2.

We now give a proof of Proposition 4.

Proof. The proof considers two separate cases: $q = 1$ and $q > 1$, since they require different methods to construct the state space.
$q = 1$: Without loss of generality we consider $p = 2$. If $\mathbf{X}_t = (\lambda_t, \lambda_{t+1})$, then $\mathbf{X}_t$ is a Markov chain. Here $\lambda_t \geq \lambda^* = \delta/(1 - \alpha_1 - \alpha_2)$. $\mathbf{X}_t$ can be constructed by iteratively imposing the random function f_u, $u \in (0, 1)$,

$$f_u : [\lambda^*, \infty) \times [\lambda^*, \infty) \longrightarrow [\lambda^*, \infty) \times [\lambda^*, \infty),$$
$$\mathbf{x} = (\lambda_1, \lambda_2) \longmapsto (\lambda_2, \delta + \alpha_1\lambda_2 + \alpha_2\lambda_1 + \beta F^{-1}_{\lambda_2}(u)).$$

For any $\mathbf{x} = (x_1, x_2), \mathbf{y} = (y_1, y_2)$ in the state space $S = [\lambda^*, \infty) \times [\lambda^*, \infty)$, let $\rho(\mathbf{x}, \mathbf{y}) = w_1|x_1 - y_1| + w_2|x_2 - y_2|$, where $w_i > 0, i = 1, 2$, and w_1, w_2 are to be decided. If $\mathbf{x}_1 = (\lambda^0_1, \lambda^0_2) := (\lambda^*, \lambda^*)$ then, for any $\mathbf{x} = (\lambda_1, \lambda_2)$, we have

$$\begin{aligned} \mathrm{E}\rho(\mathbf{X}_1(\mathbf{x}), \mathbf{X}_1(\mathbf{x}_1)) &= \int_0^1 \rho(f_u(\mathbf{x}), f_u(\mathbf{x}_1))du \\ &= a_2w_2|\lambda_1 - \lambda^0_1| + \{w_1 + w_2(a_1 + b)\}|\lambda_2 - \lambda^0_2|, \end{aligned}$$

where the last equality holds because $\lambda_t \geq \lambda^*$. Therefore it is sufficient to find an $r \in (0, 1)$ and strictly positive (w_1, w_2) such that $\mathrm{E}\rho(\mathbf{X}_1(\mathbf{x}), \mathbf{X}_1(\mathbf{x}_1)) \leq r\rho(\mathbf{x}, \mathbf{x}_1) = r\{w_1|\lambda_1 - \lambda^0_1| + w_2|\lambda_2 - \lambda^0_2|\}$. This can be obtained if the equation $r^2 - (a_1 + b)r - a_2 = 0$ yields a root $r_+ = (a_1 + b + \sqrt{(a_1 + b)^2 + 4a_2})/2 < 1$. It can be shown that under $\alpha_1 + \alpha_2 + \beta < 1$ the root $r_+ \in (0, 1)$. Here the choice of (w_1, w_2) is not unique.

$q > 1$: Without loss of generality we consider the INGARCH(2,2) model. With the Markov chain $\mathbf{X}_t = (Y_t, \lambda_t, \lambda_{t+1})$, then the chain can be obtained by defining the iterated random functions $f_u : \mathbb{Z}_0 \times [\lambda^*, \infty) \times [\lambda^*, \infty) \to \mathbb{Z}_0 \times [\lambda^*, \infty) \times [\lambda^*, \infty)$ as $f(\mathbf{x}) = f(n, \lambda_1, \lambda_2) = (F^{-1}_{\lambda_2}(u), \lambda_2, \delta + \alpha_1\lambda_2 + \alpha_2\lambda_1 + \beta_1 F^{-1}_{\lambda_2}(u) + \beta_2 n)$, where $\lambda^* = \delta/(1-\alpha_1-\alpha_2)$ and $u \in (0, 1)$. Take a metric ρ on $S = \mathbb{Z}_0 \times [\lambda^*, \infty) \times [\lambda^*, \infty)$ as $\rho(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^3 w_i|x_i - y_i|$, where $\mathbf{x} = (x_i)_{i=1}^3, \mathbf{y} = (y_i)_{i=1}^3$ and $w_i > 0, i = 1, 2, 3$. If $\mathbf{x}_1 = (n_0, \lambda^0_1, \lambda^0_2) := (0, \lambda^*, \lambda^*)$ then, for any $\mathbf{x} = (n, \lambda_1, \lambda_2)$, we have

$$\begin{aligned} \mathrm{E}\rho(\mathbf{X}_1(\mathbf{x}), \mathbf{X}_1(\mathbf{x}_1)) &= \int_0^1 |f_u(\mathbf{x}) - f_u(\mathbf{x}_1)|du \\ &= \beta_2w_3|n - n^0| + w_3\alpha_2|\lambda_1 - \lambda^0_1| \\ &\quad + \{w_1 + w_2 + (\alpha_1 + \beta_1)w_3\}|\lambda_2 - \lambda^0_2|. \end{aligned}$$

As in the first case, one needs to solve the inequality $(\alpha_2 + \beta_2)(w_1 + w_2) \leq [r - (\alpha_1 + \beta_1)](\alpha_2 + \beta_2)w_3 \leq r(w_1 + w_2)[r - (\alpha_1 + \beta_1)]$ for an $r \in (0, 1)$ and a strictly positive triple (w_1, w_2, w_3). This can be achieved if $\alpha_1 + \alpha_2 + \beta_1 + \beta_2 < 1$, which implies the quadratic equation $r^2 - (\alpha_1 + \beta_1)r - (\alpha_2 + \beta_2) = 0$ has a root $r_+ \in (0, 1)$. The result hence follows by a simple induction.

B.4. More on the NB-INGARCH

The NB-INGARCH(1, 1) model is

$$Y_t|\mathcal{F}_{t-1} \sim \text{NB}(r, p_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1},$$

where $X_t = r(1-p_t)/p_t$ and $\delta > 0, \alpha, \beta \geq 0$ are parameters. If $\alpha + \beta < 1$, then $\{X_t, t \geq 1\}$ is a geometric moment contracting Markov chain with a unique stationary distribution π and, when $X_1 \sim \pi$, $\{(X_t, Y_t), t \geq 1\}$ is ergodic. Moreover, if we assume r is known and $(\alpha+\beta)^2 + \beta^2/r < 1$, then under (L0), the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent and asymptotically normal with mean θ_0 and covariance matrix Ω^{-1}/n, where $\Omega = \text{E}\{r/X_t/(X_t + r)(\partial X_t/\partial\theta)(\partial X_t/\partial\theta)^T\}$. Verification of (L1) and (L2) is sufficient for this. Since $B^{-1}(x) = \log\{x/(x+r)\} < 0$, so (L1) holds according to Remark 1. Now $A''(\eta_t) = re^{\eta_t}/(1-e^{\eta_t})^2$ is increasing, so (L2) holds if $\gamma_Y(0) < \infty$ according to Remark 2. Because $\text{Var}(X_1) = \alpha^2\text{Var}(X_0) + \beta^2\text{Var}(Y_0) + 2\alpha\beta\text{Cov}(X_0, Y_0)$, where

$$\begin{aligned}\text{Var}(Y_0) &= \text{E}\{\text{Var}(Y_0|X_0)\} + \text{Var}\{\text{E}(Y_0|X_0)\} \\ &= \text{E}\{\frac{r(1-p_0)}{p_0^2}\} + \text{Var}(X_0) = \mu + \frac{1}{r}\text{E}X_0^2 + \text{Var}(X_0),\end{aligned}$$

and $\text{Cov}(X_1, Y_1) = \text{E}Y_1X_1 - \mu^2 = \text{E}X_1^2 - \mu^2 = \text{Var}(X_1)$, it follows from the stationarity that

$$\text{Var}(X_0) = \frac{\beta^2\mu(1+\mu/r)}{1-(\alpha+\beta)^2 - \beta^2/r}.$$

Hence $\gamma_Y(0) < \infty$ provided $(\alpha+\beta)^2 + \beta^2/r < 1$.

Appendix C. Proofs

C.1. Proof of Proposition 1

It suffices to verify the two conditions formulated in Wu and Shao (2004). For any y_0 in the state space S, $\text{E}|y_0 - f_u(y_0)| = \int_0^1 |y_0 - g(y_0, F_{y_0}^{-1}(u))|du \leq y_0 + g(0,0) + ay_0 + b\int_0^1 F_{y_0}^{-1}(u)du \leq g(0,0) + (1+a+b)y_0 < \infty$. Next, for a fixed $x_0 \in S$, there exists a unique η_0 such that $x_0 = B(\eta_0)$ due to the strict monotonicity of $B(\eta)$. For any $x \geq x_0$, there exists a unique $\eta \geq \eta_0$ such that $x = B(\eta) \geq B(\eta_0) = x_0$. Hence, by (2.3), we have

$$\begin{aligned}\text{E}|X_1(x) - X_1(x_0)| &= \int_0^1 \big|g\big(x, F_x^{-1}(u)\big) - g\big(x_0, F_{x_0}^{-1}(u)\big)\big|du \\ &\leq a|x - x_0| + b\int_0^1 \big|F_x^{-1}(u) - F_{x_0}^{-1}(u)\big|du. \qquad \text{(C.1)}\end{aligned}$$

It follows from $x \geq x_0$ and Proposition A.1 that for any $u \in (0,1)$, $F_{x_0}^{-1}(u) \leq F_x^{-1}(u)$. Therefore

$$\begin{aligned}\mathrm{E}|X_1(x) - X_1(x_0)| &\le a(x - x_0) + b\Big\{ \int_0^1 F_x^{-1}(u)du - \int_0^1 F_{x_0}^{-1}(u)du \Big\} \\ &= (a+b)(x - x_0).\end{aligned}$$

Similarly for $x < x_0$, we have $\mathrm{E}|X_1(x) - X_1(x_0)| \le (a+b)(x_0 - x)$. Now suppose $\mathrm{E}|X_n(x) - X_n(x_0)| \le (a+b)^n|x - x_0|$, then

$$\begin{aligned}\mathrm{E}|X_{n+1}(x) - X_{n+1}(x_0)| &= \mathrm{E}[\mathrm{E}\{|X_{n+1}(X_n(x)) - X_{n+1}(X_n(x_0))|\big|U_1, \ldots, U_n\}] \\ &\le \mathrm{E}\{(a+b)|X_n(x) - X_n(x_0)|\} \\ &\le (a+b)^{n+1}|x - x_0|.\end{aligned}$$

By induction, $\{X_t\}$ is geometric moment contracting and as a result, π is its unique stationary distribution.

To show that $\mathrm{E}_\pi X_1 < \infty$, by taking conditional expectation on both sides of (2.4) we have $\mathrm{E}(X_t|X_{t-1}) \le g(0,0) + (a+b)X_{t-1}$. Inductively one can show that, for any $t \ge 1$,

$$\mathrm{E}(X_t|X_1) \le \frac{1 - (a+b)^{t-1}}{1 - (a+b)} g(0,0) + (a+b)^{t-1} X_1.$$

Since for any $x \in S$, $X_t(x) \xrightarrow{\mathcal{L}} X_1 \sim \pi$ as $t \to \infty$, $X_t(0) \xrightarrow{\mathcal{L}} X_1 \sim \pi$ and, by Theorem 3.4 in Billingsley (1999), we have

$$\mathrm{E}_\pi X_1 \le \liminf_{t\to\infty} \mathrm{E}(X_t|X_1 = 0) \le \frac{g(0,0)}{1 - (a+b)} < \infty.$$

To prove (c), let $\{\xi_t, t \ge 1\}$ be a sequence of independent uniform $(0,1)$ random variables, independent of $\{X_t, t \ge 1\}$, so $Y_t = F_{X_t}^{-1}(\xi_t)$. Since $\{(X_t, \xi_t), t \ge 1\}$ is a stationary sequence if $X_1 \sim \pi$, so $\{Y_t, t \ge 1\}$ must also be a stationary process.

C.2. Proof of Proposition 2

Define a sequence of functions $\{g_k, k \ge 1\}$ in a way such that $g_1 = g$, and for $k \ge 2$, $g_k(x, y_1, \ldots, y_k) = g_{k-1}(g(x, y_k), y_1, \ldots, y_{k-1})$. Then it follows from (2.2) that for all $t \in \mathbb{Z}$,

$$X_t = g_k(X_{t-k}, Y_{t-1}, \ldots, Y_{t-k}).$$

By virtue of (2.3), we have $\mathrm{E}\big|X_t - g_1(0, Y_{t-1})\big| = \mathrm{E}\big|g_1(X_{t-1}, Y_{t-1}) - g_1(0, Y_{t-1})\big| \le a\mathrm{E}X_{t-1}$. By induction, it follows that for any $k \ge 1$,

$$\mathrm{E}\big|X_t - g_k(0, Y_{t-1}, \ldots, Y_{t-k})\big| \le a^k \ \mathrm{E}X_{t-k}.$$

Since $\mathrm{E}_\pi X_1 < \infty$, it follows that $g_k(0, Y_{t-1}, \ldots, Y_{t-k}) \xrightarrow{L^1} X_t$, as $k \to \infty$. Hence there exists a measurable function $g_\infty : \mathbb{N}_0^\infty = \{(n_1, n_2, \ldots), n_i \in \mathbb{N}_0\} \longrightarrow [0, \infty)$ such that $X_t = g_\infty(Y_{t-1}, Y_{t-2}, \ldots)$ almost surely, which proves (a).

To prove (b), let $\mathcal{F}_{k,l}^Y = \sigma\{Y_k, \ldots, Y_l\}$ for $-\infty \le k \le l \le \infty$. Then the coefficients of absolute regularity of the stationary count process $\{Y_t, t \in \mathbb{Z}\}$ are defined as

$$\beta(n) = \mathrm{E}\big\{ \sup_{A \in \mathcal{F}_{n,\infty}^Y} \big|P(A|\mathcal{F}_{-\infty,0}^Y) - P(A)\big|\big\},$$

where $\mathcal{F}_{-\infty,0}^Y = \sigma\{X_1, Y_0, Y_{-1}, \ldots\}$ according to (a). Because the distribution of $(Y_n, Y_{n+1}, \ldots)$ given $\sigma\{X_1, Y_0, Y_{-1}, \ldots\}$ is the same as that of $(Y_n, Y_{n+1}, \ldots)$ given X_1 for $n \ge 1$, the coefficients of absolute regularity become

$$\begin{aligned}\beta(n) &= \mathrm{E}\big\{ \sup_{A \in \mathcal{F}_{n,\infty}^Y} \big|P(A|\sigma\{X_1, Y_0, Y_{-1}, \ldots\}) - P(A)\big|\big\} \\ &= \mathrm{E}\big\{ \sup_{A \in \mathcal{F}_{n,\infty}^Y} \big|P(A|X_1) - P(A)\big|\big\}. \end{aligned} \tag{C.2}$$

If $\mathcal{B}^\infty$ is the σ-field in $\mathbb{R}^\infty$ generated by the cylinder sets, then we can rewrite the coefficients of absolute regularity as

$$\beta(n) = \mathrm{E}\Big\{ \sup_{A \in \mathcal{B}^\infty} \big|P\big((Y_n, Y_{n+1}, \ldots) \in A|X_1\big) - P\big((Y_n, Y_{n+1}, \ldots) \in A\big)\big|\Big\}. \tag{C.3}$$

We provide an upper bound for (C.3) by coupling the two chains $\{(X_n', Y_n'), n \in \mathbb{Z}\}$ and $\{(X_n'', Y_n''), n \in \mathbb{Z}\}$ defined on a common probability space. Suppose $X_1' \sim \pi$, $X_1'' \sim \pi$ and that X_1' is independent of X_1''. Let $\{U_k, k \in \mathbb{Z}\}$ as be an iid sequence of uniform $(0,1)$ random variables, and construct the chains as

$$\begin{aligned} X_n' &= g\big(X_{n-1}', F_{X_{n-1}'}^{-1}(U_{n-1})\big), \quad Y_n' = F_{X_n'}^{-1}(U_n), \\ X_n'' &= g\big(X_{n-1}'', F_{X_{n-1}''}^{-1}(U_{n-1})\big), \quad Y_n'' = F_{X_n''}^{-1}(U_n). \end{aligned}$$

Since X_1' and X_1'' are independent, for any $A \in \mathcal{B}^\infty$, $P((Y_n'', Y_{n+1}'', \ldots) \in A|X_1') = P((Y_n, Y_{n+1}, \ldots) \in A)$. Hence we have

$$\begin{aligned} &\big|P\big((Y_n, Y_{n+1}, \ldots) \in A|X_1 = x\big) - P\big((Y_n, Y_{n+1}, \ldots) \in A\big)\big| \\ &\quad = \big|P\big((Y_n', Y_{n+1}', \ldots) \in A|X_1' = x\big) - P\big((Y_n'', Y_{n+1}'', \ldots) \in A|X_1' = x\big)\big| \\ &\quad \le P\big((Y_n', Y_{n+1}', \ldots) \ne (Y_n'', Y_{n+1}'', \ldots)|X_1' = x\big). \end{aligned} \tag{C.4}$$

Therefore the coefficients of absolute regularity are bounded by

$$\beta(n) \le P\big((Y_n', Y_{n+1}', \ldots) \ne (Y_n'', Y_{n+1}'', \ldots)\big) \le \sum_{k=0}^{\infty} P(Y_{n+k}' \ne Y_{n+k}''). \tag{C.5}$$

The construction of the two chains agrees with the definition of geometric moment contraction (Definition 1 in Wu and Shao (2004)), so it follows from Proposition 1 that $\mathrm{E}|X_n' - X_n''| \le (a+b)^n$ for all n. Then

$$P(Y_n' \ne Y_n'') = \mathrm{E}\{P(Y_n' \ne Y_n''|X_n, X_n'')\} = \mathrm{E}\{P(|Y_n' - Y_n''| \ge 1|X_n, X_n'')\}$$

$$\leq \mathrm{E}\{E|Y_n' - Y_n''||X_n', X_n'')\} = \mathrm{E}|X_n' - X_n''| \leq (a+b)^n.$$

Hence according to (C.5), the coefficients of absolute regularity satisfy $\beta(n) \leq \sum_{k=0}^{\infty}(a+b)^{n+k} = (a+b)^n/(1-(a+b))$. Recall that β-mixing implies strong mixing (e.g., Doukhan (1994)), so $\{Y_t, t \geq 1\}$ is stationary and strongly mixing at geometric rate, in fact, it is ergodic. In particular, $\{Y_t, t \geq 1\}$ is an ergodic stationary process. It follows from $X_t = g_\infty(Y_{t-1}, Y_{t-2}, \ldots)$ that $\{X_t, t \geq 1\}$ is also ergodic.

C.3. Proof of Theorem 1

We first show identifiability and then establish the consistency result using Lemma 1. Throughout the proof, we assume that the process $\{(Y_t, X_t), t \in \mathbb{Z}\}$ is in its stationary regime. By assumption (A1), $X_t(\theta) \geq x_\theta^* \in \mathcal{R}(B)$, which implies $\eta_t(\theta) \geq B^{-1}(x_\theta^*)$. So it follows from (A2) and (A4) that, for any $\theta \in \Theta$,

$$\begin{aligned}\mathrm{E}l_t(\theta) &= \mathrm{E}\big\{Y_t B^{-1}(X_t(\theta)) - A\big(B^{-1}(X_t(\theta))\big)\big\} \\ &\leq \mathrm{E}\big\{Y_t \sup_{\theta\in\Theta} B^{-1}(X_t(\theta))\big\} - A((B^{-1}(x_\theta^*)) < \infty.\end{aligned}$$

This implies $\mathrm{E}l_t^+(\theta) < \infty$. If $M_n(\theta) = \sum_{t=1}^n l_t(\theta)/n$, then $M_n(\theta) \xrightarrow{a.s.} M(\theta) = \mathrm{E}\big\{Y_1\eta_1(\theta) - A(\eta_1(\theta))\big\}$ according to the extended mean ergodic theorem (see Billingsley (1995)). In order to prove identifiability, we need to show that θ_0 is the unique maximizer of $M(\theta)$, or that, for any $\theta \in \Theta \setminus \{\theta_0\}$, $M(\theta) - M(\theta_0) < 0$. It follows from (A5) that, for any $\theta \neq \theta_0$ and all t, $P_{\theta_0}(G_t(\theta, \theta_0)) > 0$, where $G_t(\theta, \theta_0) = \{X_t(\theta) \neq X_t(\theta_0)\}$. If $G = G_t(\theta, \theta_0)$, then we have

$$\begin{aligned}M(\theta) - M(\theta_0) &= \mathrm{E}\big[Y_t\big\{B^{-1}(X_t(\theta)) - B^{-1}\big(X_t(\theta_0)\big)\big\} \\ &\quad -\big\{A(B^{-1}(X_t(\theta))) - A(B^{-1}(X_t(\theta_0)))\big\}\big] \\ &= \mathrm{E}\big[X_t(\theta_0)\big\{B^{-1}(X_t(\theta)) - B^{-1}\big(X_t(\theta_0)\big)\big\} \\ &\quad -\big\{A(B^{-1}(X_t(\theta))) - A(B^{-1}(X_t(\theta_0)))\big\}\big] \\ &= \int_G X_t(\theta_0)\big\{B^{-1}(X_t(\theta)) - B^{-1}\big(X_t(\theta_0)\big)\big\} \\ &\quad -\big\{A(B^{-1}(X_t(\theta))) - A(B^{-1}(X_t(\theta_0)))\big\}dP_{\theta_0}.\end{aligned}$$

On the set G, there exists $c \in \mathbb{R}$ between $B^{-1}\big(X_t(\theta)\big)$ and $B^{-1}\big(X_t(\theta_0)\big)$ such that $A(B^{-1}(X_t(\theta))) - A(B^{-1}(X_t(\theta_0))) = B(c)\{B^{-1}(X_t(\theta)) - B^{-1}(X_t(\theta_0))\}$ by the Mean Value Theorem. As $A''(\eta) > 0$, $A(\eta)$ is strictly convex and c must be strictly between $B^{-1}(X_t(\theta))$ and $B^{-1}(X_t(\theta_0))$. So there exists $\xi \in \mathbb{R}$ lying strictly between $X_t(\theta)$ and $X_t(\theta_0)$ such that $\xi = B(c)$. Therefore

$$M(\theta) - M(\theta_0) = \int_G (X_t(\theta_0) - \xi)\{B^{-1}(X_t(\theta)) - B^{-1}(X_t(\theta_0))\}dP_{\theta_0}.$$

Since $B(\eta)$ is strictly increasing, $(X_t(\theta_0) - \xi)\{B^{-1}(X_t(\theta)) - B^{-1}(X_t(\theta_0))\} < 0$ if either $X_t(\theta) < X_t(\theta_0)$ or $X_t(\theta) > X_t(\theta_0)$. Hence $M(\theta) - M(\theta_0) < 0$, for any $\theta \neq \theta_0$, which establishes identifiability. To show consistency, by (A4), we have

$$\begin{aligned}\mathrm{E}\sup_{\theta\in\Theta} l_t(\theta) &= \mathrm{E}\{Y_t \sup_{\theta\in\Theta} B^{-1}(X_t(\theta)) - \inf_{\theta\in\Theta} A(B^{-1}(X_t(\theta)))\} \\ &\le \mathrm{E}\{Y_t \sup_{\theta\in\Theta} B^{-1}(X_t(\theta))\} - A(B^{-1}(x^*)) < \infty.\end{aligned}$$

The function f_θ in Lemma 1 can be taken as $f_\theta(\mathbf{y}) = y_1 B^{-1}(g^\theta_\infty(y_0, y_{-1}, \ldots)) - A(B^{-1}(g^\theta_\infty(y_0, y_{-1}, \ldots)))$, where $\mathbf{y} = (y_1, y_0, y_{-1}, \ldots)$. Hence, from (A2) and Lemma 1, $M(\theta)$ is upper-semicontinuous, and for any compact subset $K \subset \Theta$, $\limsup_{n\to\infty} \sup_{\theta\in K} M_n(\theta) \le \sup_{\theta\in K} M(\theta)$. If $\mathcal{U}_0$ is a local base of θ_0 and $U \in \mathcal{U}_0$ is a neighborhood of θ_0, then Lemma 1 can be applied to $\Theta \setminus U$. Because a u.s.c function attains its maximum on compact sets and $M(\theta) < M(\theta_0)$ for any $\theta \neq \theta_0$, we have

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta\setminus U} M_n(\theta) \le \sup_{\theta\in\Theta\setminus U} M(\theta) < M(\theta_0), \quad P_{\theta_0}\text{-a.s.} \tag{C.6}$$

Here for any $\tilde{\theta} \notin U$, $M_n(\tilde{\theta}) \le \sup_{\theta\in\Theta\setminus U} M_n(\theta)$. Let $\omega \in \Omega$ be such that (C.6) holds and $M(\theta_0) = \lim_{n\to\infty} M_n(\theta_0)$. For such ω, suppose $\hat{\theta}_n \notin U$ infinitely often, say, along a sequence denoted by $\widetilde{\mathbb{N}}$, then

$$\begin{aligned}\liminf_{n\to\infty} M_n(\hat{\theta}_n) &\le \liminf_{n\to\infty, n\in\widetilde{\mathbb{N}}} M_n(\hat{\theta}_n) \le \limsup_{n\to\infty, n\in\widetilde{\mathbb{N}}} M_n(\hat{\theta}_n) \\ &\le \limsup_{n\to\infty, n\in\widetilde{\mathbb{N}}} \sup_{\theta\notin U} M_n(\theta) \le \limsup_{n\to\infty} \sup_{\theta\notin U} M_n(\theta).\end{aligned} \tag{C.7}$$

However, according to (C.6), we have

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta\setminus U} M_n(\theta) \le \sup_{\theta\in\Theta\setminus U} M(\theta) < M(\theta_0) = \lim_{n\to\infty} M_n(\theta_0) \le \liminf_{n\to\infty} M_n(\hat{\theta}_n),$$

a contradiction. Hence there exists a null-set N_U such that for all $\omega \notin N_U$, $\hat{\theta}_n \in U$ for all n large enough. It follows by taking any set $U \in \mathcal{U}_0$ that $\hat{\theta}_n$ converges to θ_0 almost surely.

C.4. Proof of Theorem 2

We define a linearized form of $\eta_t(\theta)$ as $\eta^\dagger_t(\theta) := \eta_t(\theta_0) + (\theta - \theta_0)^T \dot{\eta}_t$, and the corresponding linearized log-likelihood function of $l(\theta)$ as

$$l^\dagger(\theta) := \sum_{t=1}^n \eta^\dagger_t(\theta) Y_t - \sum_{t=1}^n A(\eta^\dagger_t(\theta)).$$

With $u = \sqrt{n}(\theta - \theta_0)$, define

$$\begin{aligned} R_n^\dagger(u) &= l^\dagger(\theta_0) - l^\dagger(\theta_0 + un^{-1/2}) \\ &= \sum_{t=1}^n Y_t\eta_t - \sum_{t=1}^n A(\eta_t) - \sum_{t=1}^n (\eta_t + u^T n^{-1/2}\dot{\eta}_t)Y_t + \sum_{t=1}^n A(\eta_t + u^T n^{-1/2}\dot{\eta}_t) \\ &= -u^T n^{-1/2} \sum_{t=1}^n Y_t\dot{\eta}_t + \sum_{t=1}^n \{A(\eta_t + u^T n^{-1/2}\dot{\eta}_t) - A(\eta_t)\} \\ &= -u^T n^{-1/2} \sum_{t=1}^n \{Y_t - B(\eta_t)\}\dot{\eta}_t \\ &\quad + \sum_{t=1}^n \{A(\eta_t + u^T n^{-1/2}\dot{\eta}_t) - A(\eta_t) - u^T n^{-1/2} B(\eta_t)\dot{\eta}_t\}. \end{aligned} \tag{C.8}$$

If $s_t = n^{-1/2}\{Y_t - B(\eta_t)\}\dot{\eta}_t$, then $\mathrm{E}(s_t|\mathcal{F}_{t-1}) = n^{-1/2}\mathrm{E}[\{Y_t - B(\eta_t)\}\dot{\eta}_t|\mathcal{F}_{t-1}] = 0$, so $\{s_t, t \geq 1\}$ is a martingale difference sequence. Here

$$\begin{aligned} \sum_{t=1}^n \mathrm{E}(s_t s_t^T|\mathcal{F}_{t-1}) &= \frac{1}{n}\sum_{t=1}^n \mathrm{E}[\{Y_t - B(\eta_t)\}^2 \dot{\eta}_t \dot{\eta}_t{}^T|\mathcal{F}_{t-1}] \\ &= \frac{1}{n}\sum_{t=1}^n B'(\eta_t)\dot{\eta}_t\dot{\eta}_t{}^T, \end{aligned}$$

which converges almost surely to Ω by the Mean Ergodic Theorem and (A7). Moreover, for any $\epsilon > 0$,

$$\begin{aligned} &\sum_{t=1}^n \mathrm{E}\{s_t s_t^T \mathbf{1}_{[|s_t| \geq \epsilon]}|\mathcal{F}_{t-1}\} \\ &= \frac{1}{n}\sum_{t=1}^n \dot{\eta}_t\dot{\eta}_t{}^T \mathrm{E}[\{Y_t - B(\eta_t)\}^2 \mathbf{1}_{[|\{Y_t - B(\eta_t)\}\dot{\eta}_t| \geq \epsilon\sqrt{n}]}|\mathcal{F}_{t-1}] \\ &\leq \frac{1}{n}\sum_{t=1}^n \dot{\eta}_t\dot{\eta}_t{}^T \mathrm{E}[\{Y_t - B(\eta_t)\}^2 \mathbf{1}_{[|\{Y_t - B(\eta_t)\}\dot{\eta}_t| \geq M]}|\mathcal{F}_{t-1}] \\ &\longrightarrow \mathrm{E}[\{Y_1 - B(\eta_1)\}^2 \dot{\eta}_1\dot{\eta}_1{}^T \mathbf{1}_{[|\{Y_t - B(\eta_t)\}\dot{\eta}_t| \geq M]}] \quad \text{as } n \to \infty \\ &\longrightarrow 0 \quad \text{as } M \to 0. \end{aligned}$$

Then it follows from the Central Limit Theorem for martingale difference sequences that

$$\sum_{t=1}^n s_t \xrightarrow{\mathcal{L}} V \sim N(0, \Omega), \quad \text{as} \quad n \to \infty,$$

where Ω is evaluated at θ_0. The other term in (C.8), by Taylor expansion, is

$$\frac{1}{2n}\sum_{t=1}^n u^T\{B'(\eta_t)\dot{\eta}_t\dot{\eta}_t{}^T\}u + \mathcal{O}_p\Big(n^{-3/2}\sum_{t=1}^n B''(\eta_t)(u^T\dot{\eta}_t)^3\Big),$$

which is of the order of $u^T\Omega u/2 + o_P(1)$. Hence $R_n^\dagger(u) \xrightarrow{\mathcal{L}} -u^TV + (1/2)u^T\Omega u$, where $V \sim N(0,\Omega)$. It then follows that $\text{argmin}_u R_n^\dagger(u) \xrightarrow{\mathcal{L}} \text{argmin}_u\{-u^TV + (1/2)u^T\Omega u\} = \Omega^{-1}V \sim N(0,\Omega^{-1})$.

In the rest of the proof, we show that the difference between $R_n(u) := l(\theta_0) - l(\theta_0 + un^{-1/2})$ and $R_n^\dagger(u)$ is negligible as n grows large. By writing $\theta = \theta_0 + un^{-1/2}$, the difference is

$$\begin{aligned} R_n^\dagger(u) - R_n(u) = &\sum_{t=1}^n \{Y_t - B(\eta_t)\}\{\eta_t(\theta) - \eta_t - u^Tn^{-1/2}\dot\eta_t\} \\ &- \sum_{t=1}^n [A(\eta_t(\theta)) - A(\eta_t + u^Tn^{-1/2}\dot\eta_t) \\ &- B(\eta_t)\{\eta_t(\theta) - \eta_t - u^Tn^{-1/2}\dot\eta_t\}]. \end{aligned} \tag{C.9}$$

By Taylor expansion, the first term in (C.9) is $1/(2n)\sum_{t=1}^n\{Y_t - B(\eta_t)\}u^T\ddot\eta_t(\theta_t^*)u$ $= 1/(2n)u^T[\sum_{t=1}^n\{Y_t - B(\eta_t)\}\ddot\eta_t + \sum_{t=1}^n\{Y_t - B(\eta_t)\}\{\ddot\eta_t(\theta_t^*) - \ddot\eta_t\}]u$, where θ_t^* lies between θ and θ_0, and $\ddot\eta_t = \partial^2\eta_t/\partial\theta\partial\theta^T$. Since

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^n\{Y_t - B(\eta_t)\}\ddot\eta_t &\xrightarrow{a.s.} \text{E}[\{Y_t - B(\eta_t)\}\ddot\eta_t] \\ &= \text{E}[\ddot\eta_t\text{E}\{Y_t - B(\eta_t)|\mathcal{F}_{t-1}\}] = 0, \end{aligned}$$

and $1/n\sum_{t=1}^n\{Y_t - B(\eta_t)\}\{\ddot\eta_t(\theta_t^*) - \ddot\eta_t\} \xrightarrow{a.s.} 0$ under the smoothness assumption, the first term in (C.9) converges to 0 uniformly on $[-K,K]$ for any $K > 0$. We now apply a Taylor expansion to each component in the second term of (C.9),

$$\begin{aligned} A(\eta_t(\theta)) &= A(\eta_t) + u^Tn^{-1/2}B(\eta_t)\dot\eta_t \\ &\quad + \frac{1}{2n}u^T\{B(\eta_t(\theta_1^*))\ddot\eta_t(\theta_1^*) + B'(\theta_1^*)\dot\eta_t(\theta_1^*)\dot\eta_t(\theta_1^*)^T\}u, \\ A(\eta_t + u^Tn^{-1/2}\dot\eta_t) &= A(\eta_t) + B(\eta_t)u^Tn^{-1/2}\dot\eta_t + \frac{1}{2n}u^TB'(c)\dot\eta_t\dot\eta_t{}^Tu, \\ \eta_t(\theta) &= \eta_t(\theta_0 + un^{-1/2}) = \eta_t + \dot\eta_tu^Tn^{-1/2} + \frac{1}{2n}u^T\ddot\eta_t(\theta_2^*)u, \end{aligned}$$

where $0 \le c \le u^Tn^{-1/2}\dot\eta_t$, and θ_1^* and θ_2^* both lie between θ_0 and θ. Here the second term in (C.9) is

$$\begin{aligned} &\sum_{t=1}^n [A(\eta_t(\theta)) - A(\eta_t + u^Tn^{-1/2}\dot\eta_t) - B(\eta_t)\{\eta_t(\theta) - \eta_t - u^Tn^{-1/2}\dot\eta_t\}] \\ &= \sum_{t=1}^n [A(\eta_t) + u^Tn^{-1/2}B(\eta_t)\dot\eta_t + \frac{1}{2n}u^T\{B(\eta_t(\theta_1^*))\ddot\eta_t(\theta_1^*) + B'(\theta_1^*)\dot\eta_t(\theta_1^*)\dot\eta_t(\theta_1^*)^T\}u \\ &\quad - A(\eta_t) - B(\eta_t)u^Tn^{-1/2}\dot\eta_t - \frac{1}{2n}u^TB'(c)\dot\eta_t\dot\eta_t{}^Tu - B(\eta_t)\frac{1}{2n}u^T\ddot\eta_t(\theta_2^*)u] \end{aligned}$$

$$= \frac{1}{2n} u^T \sum_{t=1}^{n} [\{B(\eta_t(\theta_1^*))\ddot{\eta}_t(\theta_1^*) - B(\eta_t)\ddot{\eta}_t(\theta_2^*)\} + \{B'(\theta_1^*)\dot{\eta}_t(\theta_1^*)\dot{\eta}_t(\theta_1^*)^T - B'(c)\dot{\eta}_t\dot{\eta}_t{}^T\}]u,$$

which converges to 0 on a compact set of u under smoothness assumptions. So (C.9) converges to 0 as $n \to \infty$, which implies that $\operatorname{argmin}_u R_n(u)$ and $\operatorname{argmin}_u R_n^\dagger(u)$ have the same asymptotic distribution $\operatorname{argmin}_u R_n(u) \xrightarrow{\mathcal{L}} \Omega^{-1}V \sim N(0, \Omega^{-1})$. Here $\operatorname{argmin}_u R_n(u) = \operatorname{argmax}_u\ l(\theta_0 + un^{-1/2}) = \sqrt{n}(\hat{\theta}_n - \theta_0)$, where $\hat{\theta}_n$ is the conditional maximum likelihood estimator. Hence $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega^{-1})$, as $n \to \infty$.

C.5. Proof of Theorem 3

According to Theorems 1 and 2, it is sufficient to verify (A5). If for some $t \in \mathbb{Z}$, $X_t(\theta) = X_t(\theta_0)$, P_{θ_0}-a.s, then $\delta + \alpha X_{t-1}(\theta) + \beta Y_{t-1} = \delta_0 + \alpha_0 X_{t-1}(\theta_0) + \beta_0 Y_{t-1}$. It follows from (B.1) that

$$(\beta - \beta_0)Y_{t-1} = \delta_0 - \delta + \alpha_0\Big(\frac{\delta_0}{1-\alpha_0} + \beta_0 \sum_{k=0}^{\infty} \alpha_0^k Y_{t-k-2}\Big) - \alpha\Big(\frac{\delta}{1-\alpha} + \beta \sum_{k=0}^{\infty} \alpha^k Y_{t-k-2}\Big).$$

If $\beta \neq \beta_0$, then $Y_{t-1} \in \operatorname{span}\{Y_{t-2}, Y_{t-3}, \ldots\}$, which contradicts the fact that $\operatorname{Var}(Y_{t-1}|\mathcal{F}_{t-2}) > 0$. So β must be the same as β_0. Similarly one can show that $\alpha = \alpha_0$ and $\delta = \delta_0$, which implies $\theta = \theta_0$. Hence the model is identifiable.

C.6. Proof of Theorem 4

According to Theorem 2, we need only establish the identifiability of the model. Similar to the proof of Theorem 3, one can demonstrate that if $X_t(\theta) = X_t(\theta_0), P_{\theta_0}$-a.s. for some t, where $\theta_0 = (\delta_0, \alpha_0, \beta_0, \beta_{1,0}, \ldots, \beta_{K,0})$, then

$$(\beta - \beta_0)Y_{t-1} + \sum_{k=1}^{K} (\beta_k - \beta_{k,0})(Y_{t-1} - \xi_k)^+ = \delta_0 - \delta + \alpha_0 X_{t-1}(\theta_0) - \alpha X_{t-1}(\theta) \in \sigma\{Y_{t-2}, Y_{t-3}, \ldots\}.$$

It follows that $\beta = \beta_0$ and $\beta = \beta_{k,0}, k = 1, \ldots, K$. Similarly one can show that $\delta = \delta_0$ and $\alpha = \alpha_0$, hence $\theta = \theta_0$.

References

Billingsley, P. (1995). *Probability and Measure*. 3rd edition. Wiley, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*. 2nd edition. Wiley, New York.

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

Brockwell, A. E. (2007). Universal residuals: A multivariate transformation. *Statist. Probab. Lett.* **77**, 1473-1478.

Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*. 2nd edition. Springer.

Chang, L. (2010). Conditional modeling and conditional inference. Ph.D. thesis, Brown University.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scand. J. Statist.* **8**, 93-115.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data. *Biometrics* **65**, 1254-1261.

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

Davis, R., Dunsmuir, W. and Wang, Y. (2000). On autocorrelation in a Poisson regression models. *Biometrika* **87**, 491-506.

Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Rev.* **41**, 45-76.

Doukhan, P. (1994). *Mixing: Properties and Examples*. Springer-Verlag.

Doukhan, P., Fokianos, K. and Tjøstheim, D. (2012). On weak dependence conditions for Poisson autoregressions. *Statist. Probab. Lett.* **82**, 942-948.

Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH process. *J. Time Series Anal.* **27**, 923-942.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression. *J. Amer. Statist. Assoc.* **104**, 1430-1439.

Jung, R. and Tremayne, A. (2011). Useful models for time series of counts or simply wrong ones? *AStA Adv. Statist. Anal.* **95**, 59-91.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. 2nd edition. Springer-Verlag.

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

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrica* **14**, 249-272.

Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Samia, N. and Chan, K. (2010). Maximum likelihood estimation of a generalized threshold stochastic regression model. *Biometrika* **98**, 433-448.

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

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

Wang, W., Liu, H., Davis, R., Yao, J. F. and Li, W. K. (2014). Self-excited threshold Poisson autoregression. *J. Amer. Statist. Assoc.*.

Wu, W. and Shao, X. (2004). Limit theorems for iterated random functions. *J. Appl. Probab.* **41**, 425-436.

Yu, Y. (2009). Stochastic ordering of exponential family distributions and their mixtures. *J. Appl. Probab.* **46**, 244-254.

Zhu, F. (2010). A negative binomial integer-valued GARCH model. *J. Time Series Anal.* **32**, 54-67.

Zhu, F. (2012a). Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *J. Math. Anal. Appl.* **389**, 58-71.

Zhu, F. (2012b). Zero-inflated Poisson and negative binomial integer-valued GARCH models. *J. Statist. Plann. Inference* **142**, 826-839.

Department of Statistics, Columbia University, New York, NY 10027, USA.

E-mail: rdavis@stat.columbia.edu

Google Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA.

E-mail: vincentliu1985@gmail.com

(Received April 2014; accepted November 2014)