<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Theory and inference for a class of nonlinear models with application to time series of counts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manuscript ID</strong></td>
<td>SS-14-145tR1</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td>10.5705/ss.2014.145t</td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
<td>Richard A. Davis and Heng Liu</td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
<td>Heng Liu</td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
<td><a href="mailto:vincentliu1985@gmail.com">vincentliu1985@gmail.com</a></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
Theory and inference for a class of nonlinear models with application to time series of counts

Richard A. Davis and Heng Liu

1Department of Statistics, Columbia University
2Google 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. Moreover the 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.

Keywords: 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 et al. (2000), Davis et al. (2003), Fokianos et al. (2009), Neumann (2011), Streett (2000) and Doukhan et al. (2012). According to whether the evolution of the intensity process depends on the observations or solely on an external process, Cox (1981) classified the models into observation-driven and parameter-driven. 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 \( \lambda_t \) 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 et al. (2006) showed the second moment stationarity through a sequence of approximating processes and Fokianos et al. (2009) established the consistency and asymptotic normality of the MLE by introducing a perturbed model. However, all the above results rely heavily on the Poisson assumption and the GARCH-like dynamics of \( \lambda_t \). 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 the inference properties, such 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 et al. (2012) showed the existence of moments under similar conditions by utilizing the concept of weak dependence. More recently, Blasques et al. (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 real 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 $\sigma$-finite measure $\mu$ is given by

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

where $\eta$ is the natural parameter, and $A(\eta)$ and $h(y)$ are known functions. If $B(\eta) = A'(\eta)$, then it is known that $EY = B(\eta)$ and $\text{Var}(Y) = B'(\eta)$. The derivative of $A(\eta)$ exists generally for the exponential family, see e.g., Lehmann and Casella (1998). Since $B'(\eta) = \text{Var}(Y) > 0$, so $B(\eta)$ is strictly increasing, which establishes a one-to-one association between the values of $\eta$ and $B(\eta)$. Moreover, because we assume that the support of $Y$ is non-negative throughout this paper, so $B(\eta) = EY > 0$, 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 only the univariate case, extensions to the multi-parameter exponential family is a topic of future research.

2.2 Model formulation

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

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

(2.2)

for all $t \geq 1$, where $p(y|\eta_t)$ is defined in (2.1), $\mathcal{F}_t = \sigma\{\eta_t, Y_1, \ldots, Y_t\}$ and $X_t$ is the conditional mean process, i.e., $X_t = B(\eta_t) = E(Y_t|\mathcal{F}_{t-1})$. 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 $g_\theta$ satisfies a contraction condition, i.e., for any $x, x' \geq 0$, and $y, y' \in [0, \infty)$ or $\mathbb{N}_0$,

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

(2.3)

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

$$g_\theta(x, y) \leq g_\theta(0, 0) + ax + by, \quad \text{for any } x, y \geq 0.$$

(2.4)

We point out that model (2.2) with the function $g_\theta$ satisfying (2.3) includes the Poisson INGARCH model (see Example 1 in Section 4.1 on page 7) 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 et al. (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 $\{X_t\}$, which, by the dynamics described in (2.2), is easily seen to be a time-homogeneous Markov chain. Note that 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 turns out to be 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 $\{U_t\}_{t \geq 1}$ take values in another measurable space $\Psi$ and are independently distributed with identical marginal distribution, and $W_0$ is independent of $\{U_t\}_{t \geq 1}$.

In working with iterated random functions, Wu and Shao (2004) introduces the idea of geometric moment contraction (GMC), which is useful for deriving further properties of IRF. Our research is also relying heavily on GMC. Suppose there exists a stationary solution to the Markov chain $\{W_t\}$, denoted by $\varpi$, let $W_0, W_0' \sim \varpi$ be independent of each other and of $\{U_t\}_{t \geq 1}$, 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 $t \in \mathbb{N},$

$$E\{\rho^\alpha(W_t(W_0), W_t(W_0'))\} \leq Cr^t.$$
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 \( g_\theta \) in (2.2) evaluated at the true parameter. For any \( u \in (0,1) \), the random function \( f_u(x) \) is defined as

\[
f_u(x) := g(x, F_x^{-1}(u)),
\]

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 defined as \( Z_t(x) = (f_{U_1} \circ f_{U_2} \circ \ldots \circ f_{U_t})(x) \), which has the same distribution as \( X_t(x) \) for any \( t \).

**Proposition 1.** Assume model (2.2) and that the function \( g \) satisfies the contraction condition (2.3). Then

1. There exists a random variable \( Z_\infty \) such that, for all \( x \in S \), \( Z_n(x) \to Z_\infty \) almost surely. The limit \( Z_\infty \) does not depend on \( x \) and has distribution \( \pi \), which is the stationary distribution of \( \{X_t\} \).

2. The Markov chain \( \{X_t, t \geq 1\} \) is geometric moment contracting with \( \pi \) as its unique stationary distribution. In addition, \( E_\pi X_1 < \infty \).

3. If \( \{X_t, t \geq 1\} \) starts from \( \pi \), i.e., \( X_1 \sim \pi \), 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 \( \pi \), as \( n \to \infty \).

### 2.4 Ergodicity

In this section we further investigate the stability properties, including ergodicity and mixing for model (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 model (2.2) where the support of \( Y_t \) is a subset of \( \mathbb{N}_0 = \{0,1,\ldots\} \), and that \( g \) satisfies the contraction condition (2.3). Then

1. There exists a measurable function \( g_\infty : \mathbb{N}_0^\infty = \{(n_1,n_2,\ldots) , n_i \in \mathbb{N}_0 , i = 1,2,\ldots \} \to [0,\infty) \)

   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

   \[
   \beta(n) \leq (a + b)^n / (1 - (a + b)),
   \]

   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, including consistency and asymptotic normality. Denote the $d$-dimensional parameter vector by $\theta \in \mathbb{R}^d$, i.e., $\theta = (\theta_1, \ldots, \theta_d)^T$, and the true parameter vector by $\theta_0 = (\theta_0^1, \ldots, \theta_0^d)^T$. Then the likelihood function of model (2.2) conditioned on $\eta_1$ and based on the observations $Y_1, \ldots, Y_n$ is given by

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

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, up to a constant independent of $\theta$, is given by

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

with score function

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

The maximum likelihood estimator $\hat{\theta}_n$ is a solution to the equation $S_n(\theta) = 0$. Let $P_{\theta_0}$ be the probability measure under the true parameter $\theta_0$ and unless otherwise indicated, $E[\cdot]$ is taken under $\theta_0$. Recall that $X_t = g^0_\infty(Y_{t-1}, Y_{t-2}, \ldots)$ according to part (a) of Propositions 2. We will derive the asymptotic properties of the maximum likelihood estimator $\hat{\theta}_n$ based on a set of regularity conditions:

(A0) $\theta_0$ is an interior point in the compact parameter space $\Theta \subset \mathbb{R}^d$.

(A1) For any $\theta \in \Theta$, $g^0_\infty \geq x^*_\theta \in \mathcal{R}(B)$, where $\mathcal{R}(B)$ is the range of $B(\eta)$. Moreover $x^*_\theta \geq x^* \in \mathcal{R}(B)$ for all $\theta$.

(A2) For any $y \in [0, \infty)^d$ or $N_0^d$, the mapping $\theta \mapsto g^0_\infty(y)$ is continuous.

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

(A4) $E\{Y_t \sup_{\theta \in \Theta} B^{-1}(g^0_\infty(Y_0, Y_{-1}, \ldots))\} < \infty$.

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

(A6) The mapping $\theta \mapsto g^0_\infty$ is twice continuously differentiable.

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

Strong consistency of the estimates is derived according to the lemma below, which 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 $\{f_\theta : \mathbb{R}^d \mapsto [-\infty, \infty], \theta \in \Theta\}$ be a family of Borel measurable functions such that:

1. $\theta \mapsto f_\theta(x)$ is upper-semicontinuous for all $x \in \mathbb{R}^d$. 


2. \( \sup_{\theta \in C} f_{\theta}(x) \) is Borel measurable for any compact set \( C \subset \Theta \).

3. \( \mathbb{E}\{\sup_{\theta \in \Theta} f_{\theta}(X)\} < \infty \) for some random variable \( X \) defined on \( (\Omega, \mathcal{F}, P) \).

Then

1. \( \theta \mapsto \mathbb{E}[f_{\theta}(X)] \) 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

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

for any compact set \( C \).

Pfanzagl (1969) proved the result assuming the independent structure of \( \{X_t\} \), but the same result proves to be true provided that the strong law of large numbers can be applied. By virtue of Lemma 1, we can derive the strong consistency of the estimates.

**Theorem 1.** Assume model (2.2) with the function \( g \) satisfying the contraction condition (2.3), and that assumptions (A0)-(A5) hold. Then the maximum likelihood estimator \( \hat{\theta}_n \) is strongly consistent, that is,

\[
\hat{\theta}_n \xrightarrow{a.s.} \theta_0, \quad \text{as } n \to \infty.
\]

The following theorem addresses the asymptotic distribution of the MLE and the idea of proof is similar to that in Davis et al. (2003). Unless otherwise indicated, \( \eta_t \) and \( \dot{\eta}_t \) are both evaluated at \( \theta_0 \), i.e., \( \eta_t = \eta_t(\theta_0) \) and \( \dot{\eta}_t = (\partial \eta_t / \partial \theta)|_{\theta = \theta_0} \).

**Theorem 2.** Assume model (2.2) with the function \( g \) satisfying the contraction condition (2.3), and that assumptions (A0)-(A7) hold. Then the maximum likelihood estimator \( \hat{\theta}_n \) is asymptotically normal, i.e.,

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

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

We remark that in practice, the population quantities in \( \Omega \) can be replaced by their estimated counterparts. Examples of such substitution will be illustrated below in specific models.

4 **Examples**

4.1 **Linear dynamic models**

The conditional mean process \( \{X_t\} \) in these models has GARCH-like dynamics. Specifically they are described as

\[
Y_t|\mathcal{F}_{t-1} \sim p(y|\eta_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1}, \quad (4.1)
\]
where \( X_t = B(\eta_t) = E(Y_t|F_{t-1}) \), and \( \delta > 0, \alpha, \beta \geq 0 \) are parameters. Observe that model (4.1) is a special case of model (2.2) by defining the function \( g_\theta \) as

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

with \( \theta = (\delta, \alpha, \beta)^T \) and the contraction condition (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, denote \( d_t = Y_t - X_t \), then it follows from \( E(d_t|F_{t-1}) = 0 \) that \( \{d_t, t \in \mathbb{Z}\} \) is a martingale difference sequence. Therefore model (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 model (4.1).

**Proposition 3.** Assume model (4.1) with \( \alpha + \beta < 1 \). Then the process \( \{X_t, t \geq 1\} \) has a unique stationary distribution \( \pi \), 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 model (4.1) are given by (3.1) and (3.2) respectively, where \( \partial \eta_l(\theta)/\partial \theta = (\partial \eta_l/\partial \delta, \partial \eta_l/\partial \alpha, \partial \eta_l/\partial \beta)^T \) is determined recursively by

\[
  \frac{\partial \eta_l}{\partial \theta} = \left( \frac{1}{B(\eta_{l-1})} / B'(\eta_l) + \alpha \frac{B'(\eta_{l-1})}{B'(\eta_l)} \frac{\partial \eta_{l-1}}{\partial \theta} \right). \tag{4.4}
\]

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

- (L0) The true parameter vector \( \theta_0 \) lies in a compact neighborhood \( \Theta \subset \mathbb{R}_+^3 \) of \( \theta_0 \), 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) \( E\{Y_1 \sup_{\theta \in \Theta} B^{-1}(\delta/(1-\alpha) + \beta \sum_{k=0}^{\infty} \alpha^k Y_{-k})\} < \infty \).
- (L2) \( E\{B'(\eta_l(\theta_0))(\partial \eta_l(\theta)/\partial \theta_i)^2|\theta = \theta_0\} < \infty \), for \( i = 1, 2, 3 \).

**Theorem 3.** Assume model (4.1) and that assumptions (L0)-(L2) hold. Then the maximum likelihood estimator \( \hat{\theta}_n \) is strongly consistent and asymptotically normal, i.e.,

\[
  \sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow N(0, \Omega^{-1}), \quad \text{as} \quad n \to \infty,
\]

where \( \Omega = E\{B'(\eta_l)\eta_l\eta_l^T\} \), where \( \eta_l = \eta_l(\theta_0) \) and \( \eta_l = \frac{\partial \eta_l}{\partial \theta}|_{\theta = \theta_0} \).

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

**Remark 1.** A sufficient condition for assumption (L1) is

\[
  E\{Y_1 B^{-1}(\delta_U/\epsilon + \sum_{k=1}^{\infty} (1-\epsilon)^k Y_{1-k})\} < \infty, \quad \text{and}
\]

\[
  |F_\theta| \leq 1 - \epsilon, \quad \text{for some} \quad \epsilon > 0.
\]
provided that $\delta_U/\epsilon + \sum_{k=1}^{\infty} (1-\epsilon)^k Y_{1-k}$ is in the range of $B(\eta)$.

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

Next we consider two examples as special cases of model (4.1).

Example 1. The Poisson INGARCH(1,1) model is given by

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

where $\delta > 0$, $\alpha, \beta \geq 0$ are parameters. If $\alpha + \beta < 1$, then $\{\lambda_t\}$ has a unique stationary distribution $\pi$; 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.

Moreover, 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

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

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 stationarity result.

Proposition 4. Consider the INGARCH($p,q$) model (4.6) and suppose $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$, 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 defined as

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

where $X_t = r(1-p_t)/p_t$, $\delta > 0$, $\alpha, \beta \geq 0$ are parameters and the notation $Y \sim \text{NB}(r,p)$ represents the negative binomial distribution with probability mass function given by

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

When $r = 1$, the conditional distribution of $Y_t$ becomes geometric distribution 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 $\pi$; 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 thus $\hat{\theta}$ can be obtained correspondingly. Moreover, if we assume $r$ is known and $(\alpha + \beta)^2 + \beta^2/r < 1$, then under the assumption (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 et al. (2003). In this framework, the model specification is given by

\[
Y_t | F_{t-1} \sim p(y | \eta_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1} + \sum_{k=1}^{K} \beta_k (Y_{t-1} - \xi_k)^{+},
\]  

(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 model (4.8) is a special case of model (2.2) by defining \( g_0(x, y) = \delta + \alpha x + \beta y + \sum_{k=1}^{K} \beta_k (y - \xi_k)^{+} \), where \( \theta = (\delta, \alpha, \beta, \beta_1, \ldots, \beta_K)^T \).

Being piecewise linear in \( Y_{t-1} \), model (4.8) in principle provides a wide and flexible family of nonlinear dynamics. Note that 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) and an attempt to model negative autocorrelation. Our model (4.8) turns out to be a special case of it, in that the same \( \alpha \) is shared across regimes. According to Propositions 1 and 2, we can establish the stability properties of model (4.8).

**Proposition 5.** Consider model (4.8) with parameters satisfying \( \alpha + \beta < 1, \beta + \sum_{k=1}^{s} \beta_k \leq 0 \) 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 \( \pi \). Moreover if \( X_1 \sim \pi \), then \( \{(X_t, Y_t), t \geq 1\} \) is ergodic.

We now consider inference for this model. 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 employ both procedures to real datasets in the next section.

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

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

(4.9)

This defines the function \( g_\infty^\theta \) as in \( X_t = g_\infty^\theta (Y_{t-1}, Y_{t-2}, \ldots) \) and also verifies assumptions (A1)-(A3). Hence in order to apply Theorem 3, we only need to impose the following regularity assumptions for the nonlinear model (4.8):

(NL1) \( \theta_0 \) is an interior point in the parameter space \( \Theta \), which is a compact subset of the parameter set satisfying the conditions in Proposition 5.

(NL1) \( \mathbb{E}[\sup_{\theta \in \Theta} B^{-1}(\frac{\delta}{(1 - \alpha)} + \sum_{i=0}^{\infty} \alpha^i \{\beta Y_{t-1-i} + \sum_{k=1}^{K} \beta_k (Y_{t-1-i} - \xi_k)^{+}\}] < \infty. \)
Sufficient conditions for assumptions (NL1) and (NL2) can be established similarly to those given in Remarks 1 and 2 in Appendix B. The asymptotic properties of the MLE are summarized in the following theorem.

Theorem 4. For model (4.8), suppose that the placement of the knots is known, and that assumptions (NL0)-(NL2) hold, then the maximum likelihood estimator \( \hat{\theta}_n \) is strongly consistent and asymptotically normal, i.e.,

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

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

We use the Poisson nonlinear dynamic model as an illustrative example of the above results and refer readers to Section 5 for implementation of the estimation procedure. The model is defined as

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

(4.10)

It follows that under the conditions of Proposition 5 and Theorem 4 that \( \{(\lambda_t, Y_t), t \geq 1\} \) 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

\[
\frac{\partial \lambda_t}{\partial \theta} = (1 \quad \lambda_{t-1} \quad (Y_{t-1} - \xi_1)^+ \quad \ldots \quad (Y_{t-1} - \xi_K)^+)^T + \alpha \frac{\partial \lambda_{t-1}}{\partial \theta}.
\]

Another example of nonlinear dynamic models is the Poisson exponential autoregressive model proposed by Fokianos et al. (2009), and it is given by

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

(4.11)

where \( \alpha_0, \alpha_1, \beta, \gamma > 0 \) are parameters. We point out that if \( \alpha_0 + \alpha_1 + \beta < 1 \), then model (4.11) belongs to the class of models (2.2) and hence enjoys the stability properties stated in Propositions 1 and 2. As for the inference of the model, we refer readers to Fokianos et al. (2009) for details.

5 Numerical results

The performance of the estimation procedure for the Poisson nonlinear dynamic model is illustrated in a simulation study. The MLE is 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 et al. (2009). Other models including the negative binomial linear and nonlinear dynamic models and the exponential autoregressive model (4.11) will be applied to two real datasets, and tools for checking goodness of fit will be considered.

5.1 Simulation for the nonlinear model

As specified in (4.10), a 1-knot nonlinear dynamic model is simulated according to

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

with different sample sizes. Each sample size and parameter configuration is replicated 1000 times. For each realization, the first 500 simulated observations are discarded as burn-in in order to let the
Table 1: Estimation results for 1-knot model with known knot location

<table>
<thead>
<tr>
<th></th>
<th>δ</th>
<th>α</th>
<th>β</th>
<th>β₁</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>Estimates</td>
<td>0.5596</td>
<td>0.4861</td>
<td>0.3990</td>
<td>-0.2009</td>
<td>500</td>
</tr>
<tr>
<td>s.e.</td>
<td>(0.0087)</td>
<td>(0.0030)</td>
<td>(0.0026)</td>
<td>(0.0051)</td>
<td></td>
</tr>
<tr>
<td>Estimates</td>
<td>0.5265</td>
<td>0.4944</td>
<td>0.3991</td>
<td>-0.2016</td>
<td>1000</td>
</tr>
<tr>
<td>s.e.</td>
<td>(0.0041)</td>
<td>(0.0016)</td>
<td>(0.0013)</td>
<td>(0.0025)</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>δ</th>
<th>α</th>
<th>β</th>
<th>β₁</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>Estimates</td>
<td>0.5387</td>
<td>0.4852</td>
<td>0.4187</td>
<td>-0.1614</td>
<td>500</td>
</tr>
<tr>
<td>s.e.</td>
<td>(0.0089)</td>
<td>(0.0030)</td>
<td>(0.0031)</td>
<td>(0.0047)</td>
<td></td>
</tr>
<tr>
<td>Estimates</td>
<td>0.5002</td>
<td>0.4943</td>
<td>0.4197</td>
<td>-0.1679</td>
<td>1000</td>
</tr>
<tr>
<td>s.e.</td>
<td>(0.0042)</td>
<td>(0.0016)</td>
<td>(0.0015)</td>
<td>(0.0023)</td>
<td></td>
</tr>
</tbody>
</table>

process reach its stationary regime. We first estimate the parameters assuming that the location of the knot is known, i.e., the true underlying model is (4.8) with only one knot at 5. The means and standard errors of the estimates from all 1000 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 described 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 1000 replications are summarized in Table 2. Here the locations of the knots are determined by sample quantiles. Not surprisingly, the performance of the maximum likelihood estimates of $\beta$ and $\beta_1$ is not as good as in the known knot case. However, 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 instance when $n = 1000$, the means of ranked probability scores (RPS) for known and unknown knot cases are 1.0906 and 1.0914, respectively.

Next we turn to the problem of selecting the number of knots using an information criterion. Simulations with different sample sizes are implemented and the model selection results are summarized in Table 3. Numbers in the table stand for the proportion of times that each particular model is selected in the 1000 runs. For AIC, the 1-knot model is selected most often followed by a 2-knot model, at least in the cases when $n = 1000$.

In light of the idea of interpolating the nonlinear dynamic of $\lambda_t$ 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 $0.4y - 0.2(y - 5)^+$. From the graph, we can see that the piecewise linear function fitted by the 1-knot model is closest to the true curve.

Table 3: Model selection of 1-knot simulation

<table>
<thead>
<tr>
<th>Criteria</th>
<th>0 knot</th>
<th>1 knot</th>
<th>2 knots</th>
<th>3 knots</th>
<th>≥ 4 knots</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>34.3%</td>
<td>37.6%</td>
<td>20.9%</td>
<td>5.2%</td>
<td>2.0%</td>
<td>500</td>
</tr>
<tr>
<td>BIC</td>
<td>80.5%</td>
<td>18.8%</td>
<td>0.6%</td>
<td>0.1%</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>12.4%</td>
<td>45.0%</td>
<td>29.9%</td>
<td>8.3%</td>
<td>4.4%</td>
<td>1000</td>
</tr>
<tr>
<td>BIC</td>
<td>59.4%</td>
<td>38.4%</td>
<td>2.0%</td>
<td>0.2%</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Histograms of the 1-knot model with sample size 1000 assuming the knot is known. The overlaying curves are the density estimates and the dashed vertical lines represent the true values of the parameters.
Figure 2: Left: the 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.
Figure 3: Top: Number of transactions per minute of the stock Ericsson B during July 2nd 2002; Bottom: ACF of the data.

5.2 Two data applications

1. Number of transactions of Ericsson stock

As an illustrative example, both linear and nonlinear dynamic models are 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.

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

\[ \hat{\lambda}_t = 0.2912 + 0.8312\hat{\lambda}_{t-1} + 0.1395Y_{t-1}, \]

\[ (0.1000) \quad (0.0242) \quad (0.0188) \]

and the fitted NB-INGARCH model is

\[ Y_t|\mathcal{F}_{t-1} \sim NB(8, \hat{p}_t), \quad \hat{X}_t = 0.2676 + 0.8447\hat{X}_{t-1} + 0.1282Y_{t-1}, \]

\[ (0.1406) \quad (0.0350) \quad (0.0274) \]

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

As for the Poisson nonlinear dynamic model, AIC and BIC are used to help select the number of knots among 0 to 5; the values are reported in Table 4. The fitted 1-knot Poisson model, which has the smallest AIC, is given by

\[ \hat{\lambda}_t = 0.5837 + 0.8319\hat{\lambda}_{t-1} + 0.0906Y_{t-1} + 0.0722(Y_{t-1} - 9)^{+}. \]

\[ (0.1884) \quad (0.0241) \quad (0.0295) \quad (0.0373) \]

Note that 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 $\hat{\lambda}_t = 0.5519 + 0.8326\hat{\lambda}_{t-1} + 0.0961Y_{t-1} + 0.0154(Y_{t-1} - 7)^+ + 0.0559(Y_{t-1} - 11)^+$ 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.

As can be seen from the model checking below, 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 select the number of knots by minimizing the AIC. It turns out that the AIC value of a 1-knot model is the second smallest among all the candidates, with 2674.69 compared to the smallest value 2674.04, which is attained by the negative binomial INGARCH model fitted above. The fitted 1-knot negative binomial nonlinear model is given by $Y_t|\mathcal{F}_{t-1} \sim \text{NB}(8, \hat{p}_t)$, where $\hat{X}_t = 8(1 - \hat{p}_t)/\hat{p}_t$ follows

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

(0.2559) (0.0350) (0.0412) (0.0546)

Here the locations of knots for the nonlinear dynamic model are all 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 are nearly identical. The exponential autoregressive model (4.11) is also applied to this dataset by Fokianos et al. (2009) and is given by

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

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

To assess the adequacy of the fit by all of the above models, we will 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 et al. (2003) and Jung and Tremayne (2011) for a comprehensive treatment of the tools. In our study, we first consider the standardized Pearson residuals $e_t = (Y_t - E(Y_t|\mathcal{F}_{t-1}))/\sqrt{\text{Var}(Y_t|\mathcal{F}_{t-1})}$ which can be obtained by replacing the population quantities by their estimated counterparts. If the model is correctly specified, then the residuals $\{e_t\}$ should be a white noise sequence with constant variance. It turns out that 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 tool for model checking is through the probability integral transform (PIT). When the underlying distribution is continuous, it is well known that the PIT follows standard uniform distribution. However, if the underlying distribution is discrete, some adjustments are required and the so-called randomized PIT is therefore introduced by perturbing the step function characteristic of the CDF of discrete random variables (see Brockwell (2007)). More recently, Czado et al. (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 defined by

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

where $\nu_t$ is a uniform random variable on $[0, 1]$. The randomized PIT is given by

$$
\tilde{u}_t = F_t(Y_t) + \nu_t - \frac{1}{2}.
$$

This transformation is known to improve the uniformity of the PIT.

Table 4: Model selection results for Ericsson data

<table>
<thead>
<tr>
<th></th>
<th>0-knot</th>
<th>1-knot</th>
<th>2-knot</th>
<th>3-knot</th>
<th>4-knot</th>
<th>5-knot</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogL</td>
<td>-1433.19</td>
<td>-1431.21</td>
<td>-1431.08</td>
<td>-1430.58</td>
<td><strong>-1429.65</strong></td>
<td>-1431.12</td>
</tr>
<tr>
<td>AIC</td>
<td>2874.38</td>
<td><strong>2872.41</strong></td>
<td>2874.17</td>
<td>2875.17</td>
<td>2875.30</td>
<td>2880.25</td>
</tr>
<tr>
<td>BIC</td>
<td><strong>2890.90</strong></td>
<td>2893.07</td>
<td>2898.95</td>
<td>2904.08</td>
<td>2908.35</td>
<td>2917.43</td>
</tr>
</tbody>
</table>

Table 4: Model selection results for Ericsson data

<table>
<thead>
<tr>
<th></th>
<th>0-knot</th>
<th>1-knot</th>
<th>2-knot</th>
<th>3-knot</th>
<th>4-knot</th>
<th>5-knot</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogL</td>
<td>-1433.19</td>
<td>-1431.21</td>
<td>-1431.08</td>
<td>-1430.58</td>
<td><strong>-1429.65</strong></td>
<td>-1431.12</td>
</tr>
<tr>
<td>AIC</td>
<td>2874.38</td>
<td><strong>2872.41</strong></td>
<td>2874.17</td>
<td>2875.17</td>
<td>2875.30</td>
<td>2880.25</td>
</tr>
<tr>
<td>BIC</td>
<td><strong>2890.90</strong></td>
<td>2893.07</td>
<td>2898.95</td>
<td>2904.08</td>
<td>2908.35</td>
<td>2917.43</td>
</tr>
</tbody>
</table>
1−knot NB model for Ericsson data

ACF of standardized Pearson residuals

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.

where \( \{ \nu_t \} \) is a sequence of iid uniform \((0, 1)\) random variables, \( F_t(\cdot) \) is the predictive cumulative distribution. In our situation, \( F_t(\cdot) \) is simply the CDF of a Poisson or a negative binomial distribution. If the model is correct, then \( \tilde{\nu}_t \) 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 follows \((0, 1)\) uniform distribution, the histograms of PIT from different models are plotted and a Kolmogorov-Smirnov test is carried out. The results are summarized in Figure 5, and the p-values are reported in Table 5. It can be seen that both of the two negative binomial-based models pass the PIT test, while none of the Poisson-based models does. This observation could be explained, as mentioned above, 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 et al. (2009) and Jung and Tremayne (2011). Most of them are computed as the average of quantities related to predictions and take the form \((n - 1)^{-1} \sum_{i=2}^{n} s(F_t(Y_i))\) where \( F_t(\cdot) \) is the CDF of the prediction distribution and \( s(\cdot) \) denotes some scoring rule. In this paper we calculate three scoring rules: logarithmic score (LS), quadratic score (QS) and ranked probability score (RPS), as a basis for evaluating the relative performance of our fitted models. For definition of these scores, 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

As a second example, we construct a time series based on daily log-returns of Goldman Sachs Group (GS) stock from May 4th, 1999 to March 16th, 2012. We first calculate 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 will be 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
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 $45^\circ$ line with zero intercept.

Table 5: Quantitative model checking for Ericsson data

<table>
<thead>
<tr>
<th>Model</th>
<th>log likelihood</th>
<th>$p$-value of PIT</th>
<th>LS</th>
<th>QS</th>
<th>RPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson INGARCH</td>
<td>-1433.19</td>
<td>$&lt; 10^{-5}$</td>
<td>3.1167</td>
<td>-0.0576</td>
<td>2.6883</td>
</tr>
<tr>
<td>NB INGARCH</td>
<td>-1332.02</td>
<td>0.7386</td>
<td>2.8958</td>
<td>-0.0671</td>
<td>2.6063</td>
</tr>
<tr>
<td>1-knot Poisson</td>
<td>-1431.21</td>
<td>$&lt; 10^{-5}$</td>
<td>3.1123</td>
<td>-0.0573</td>
<td>2.6848</td>
</tr>
<tr>
<td>2-knot Poisson</td>
<td>-1431.08</td>
<td>$&lt; 10^{-5}$</td>
<td>3.1121</td>
<td>-0.0575</td>
<td>2.6843</td>
</tr>
<tr>
<td>3-knot Poisson</td>
<td>-1430.58</td>
<td>$&lt; 10^{-5}$</td>
<td>3.1110</td>
<td>-0.0580</td>
<td>2.6779</td>
</tr>
<tr>
<td>1-knot NB model</td>
<td>$-1331.34$</td>
<td>0.8494</td>
<td>2.8942</td>
<td>$-0.0671$</td>
<td>2.6021</td>
</tr>
<tr>
<td>Exp-auto model</td>
<td>-1448.69</td>
<td>$&lt; 10^{-5}$</td>
<td>3.1504</td>
<td>-0.0600</td>
<td>2.6924</td>
</tr>
</tbody>
</table>
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 $p = 0.1$.

generically 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 return times. Note that 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, three models: the geometric INGARCH (negative binomial INGARCH (4.7) with $r = 1$), and the 1-knot and 2-knot geometric-based models are fitted to the data. The number of knots for the nonlinear dynamic models is chosen by minimizing the AIC, and the locations of knots are estimated by maximizing the likelihood based on a grid search. In addition, the following constraint is 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 it follows from the definition of return times that $Y_t \geq 1$ for any $t$, we use a version of
the geometric distribution that counts the total number of trials, instead of only the failures. In particular, the fitted 1-knot geometric-based model is given by

\[ Y_t - 1 \mid F_{t-1} \sim \text{Geom}(p_t), \]

where \( X_t = (1 - p_t) / p_t \). Notice that in both models, \( \hat{\alpha} + \hat{\beta} \) is very close to unity, i.e., 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 1101. A simple simulation according to the fitted model yields the mean and median very close to those of the data, but the variance of the simulated data is extraordinarily large, which resembles the feature of the observed data. This is because, although the fitted models are still stationary, the parameters no longer satisfy the conditions specified in Theorem 4 that ensure a finite variance.

It turns out that the geometric-based models fitted above are capable of capturing the high volatility part of the data. Their standardized Pearson residuals are 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 that is heavily used in this paper is the 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.

**Proposition 6.** Suppose two random variables \( Y' \) and \( Y'' \) follow distributions belonging to the one-parameter exponential family (2.1) with the same \( A, h \) and \( \mu \), but with natural parameters \( \eta' \) and \( \eta'' \) respectively. If \( \eta' \leq \eta'' \), then \( Y' \) is stochastically smaller than \( Y'' \).
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.
Proof. Denote the probability density functions of \( Y' \) and \( Y'' \) as \( p(y|\eta') \) and \( p(y|\eta'') \) defined in (2.1), respectively. Then the log ratio of the two densities is

\[
l(y) = \log \frac{p(y|\eta')}{p(y|\eta'')} = \log \frac{\exp\{\eta'y - A(\eta')\}h(y)}{\exp\{\eta''y - A(\eta'')\}h(y)} = y(\eta' - \eta'') + [A(\eta'') - A(\eta')],
\]

which is apparently a concave function in \( y \). So it follows from Definition 2 in Yu (2009) that \( Y' \) is log concave relative to \( Y'' \), i.e., \( Y' \leq_{tc} Y'' \). Moreover, since \( A(\eta) \) is increasing in \( \eta \), so \( \lim_{\eta \to 0} l(y) = A(\eta'') - A(\eta') \geq 0 \) 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 \( Y'' \), i.e., \( Y' \leq_{st} Y'' \). \( \square \)

Denote \( F_x \) as 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 \( u \in [0,1] \). The result below provides a useful tool for the coupling technique employed to establish mixing conditions for the observation process.

Proposition 7. Suppose that \( U \) is a uniform \((0,1)\) random variable, and define two random variables \( Y' \) and \( Y'' \) as

\[
Y' = F_x^{-1}(U) \quad \text{and} \quad Y'' = F_{x''}^{-1}(U),
\]

where \( x' = B(\eta') \) and \( x'' = B(\eta'') \). Then \( E|Y' - Y''| = |x' - x''| \).

Proof. It follows from the construction of \( Y' \) and \( Y'' \) that they follow the one-parameter exponential family (2.1) with natural parameters \( \eta' \) and \( \eta'' \) respectively, and \( EY' = x', \ EY'' = x'' \). If \( x' \leq x'' \), then \( Y' \) is stochastically smaller than \( Y'' \) by virtue of Proposition 6. It follows that \( F_x^{-1}(\theta) \leq F_{x''}^{-1}(\theta) \) for \( \theta \in (0,1) \), i.e., \( Y' \leq Y'' \). This implies \( E|Y' - Y''| = E(Y'' - Y') = x'' - x' \). Similarly if \( x' \geq x'' \), then \( E|Y' - Y''| = x' - x'' \). Hence we have \( E|Y' - Y''| = |x' - x''| \). \( \square \)

Appendix B. Linear dynamic models and examples

For linear dynamic model (4.1), note that by recursion we have, for all \( t \),

\[
X_t(\theta) = \delta/(1 - \alpha) + \beta \sum_{k=0}^{\infty} \alpha^k Y_{t-1-k}.
\]  

(5.1)

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, i.e.,

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

where

\[
\frac{\partial^2 \eta_t(\theta)}{\partial \theta \partial \theta^T} = \left( \begin{array}{c} \frac{B''(\eta_t(\theta))}{(B'(\eta_t(\theta))^2} \frac{\partial \eta_t(\theta)}{\partial \theta} \frac{B'(\eta_t(\theta))}{(B'(\eta_t(\theta)))} \frac{\partial \eta_t(\theta)}{\partial \theta} \frac{B'(\eta_t(\theta))}{(B'(\eta_t(\theta)))} \frac{\partial \eta_t(\theta)}{\partial \theta} \\
\frac{\eta_t-1}{(B'(\eta_t(\theta)))} \frac{\partial \eta_t(\theta)}{\partial \theta} \frac{B'(\eta_t(\theta))}{(B'(\eta_t(\theta)))} \frac{\partial \eta_t(\theta)}{\partial \theta} \frac{B'(\eta_t(\theta))}{(B'(\eta_t(\theta)))} \frac{\partial \eta_t(\theta)}{\partial \theta} + (0 \ 1 \ 0)^T \frac{B'(\eta_t(\theta))}{B'(\eta_t(\theta))} \frac{\partial \eta_t-1}{\partial \theta} + \alpha \frac{B''(\eta_t(\theta))}{(B'(\eta_t(\theta)))} B'(\eta_t(\theta)) \\
\frac{\partial \eta_t-1}{\partial \theta} \frac{\partial \eta_t-1}{\partial \theta^T} - \alpha \frac{B''(\eta_t-1)}{(B'(\eta_t(\theta)))} \frac{\partial \eta_t}{\partial \theta} \frac{\partial \eta_t}{\partial \theta^T} + \alpha \frac{B'(\eta_t-1)}{(B'(\eta_t(\theta)))} \frac{\partial^2 \eta_t}{\partial \theta \partial \theta^T}. \end{array} \right)
\]
\section*{B.2. Proofs of Remarks 1 and 2}

\textit{Proof.} Remark 1 can be seen by noting that \( X_1(\theta) \leq \frac{\delta_U}{\epsilon} + \sum_{k=1}^{\infty} (1 - \epsilon)^k Y_{t-k} \). 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 \( \delta \) and \( \beta \) are similar. First note that

\[
E\{B'(\eta_0(\theta_0))(\frac{\partial \eta_0(\theta_0)}{\partial \alpha})^2\} = E\{ \frac{1}{B'(\eta_0)}(\frac{\partial B(\eta_0)}{\partial \alpha})^2 \} \leq \frac{1}{\epsilon} E\{ \frac{\partial B(\eta_0)}{\partial \alpha} \}^2,
\]

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

\[
E\{ \sum_{k=1}^{\infty} k\alpha^{k-1}Y_{t-k} \}^2 \leq \{ \gamma_Y(0) + \frac{2\gamma_Y(1)}{1-\alpha(\alpha+\beta)} \} \sum_{k=1}^{\infty} k^2\alpha^{2k-2}
\]

\[
+ \frac{2\alpha \gamma_Y(1)}{1-\alpha^2(\alpha+\beta)^2} \sum_{k=1}^{\infty} k\alpha^{2k-2} + \mu^2 \sum_{k=1}^{\infty} k\alpha^{k-1} < \infty,
\]

where \( \mu = EY_1 < \infty \). Hence \( E[B'(\eta_0(\theta_0))(\partial \eta_0(\theta_0)/\partial \alpha)^2] < \infty \) if \( \gamma_Y(0) < \infty \). \hfill \Box

\section*{B.3. More on the Poisson INGARCH}

As illustrated in Example 1 in Section 4.1, the Poisson INGARCH(1, 1) model is given by

\[ Y_t|F_{t-1} \sim \text{Pois} (\lambda_t), \quad \lambda_t = \delta + \alpha \lambda_{t-1} + \beta Y_{t-1}, \]

where \( \delta > 0, \alpha, \beta \geq 0 \) are parameters. When \( \alpha + \beta < 1 \), \( \{(Y_t, \lambda_t), t \geq 1\} \) is an ergodic stationary process; in addition, the MLE \( \hat{\theta}_n \) is strongly consistent and asymptotically normal according to Theorem 3, i.e., \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{L}{\to} N(0, \Omega^{-1}) \), as \( n \to \infty \), where \( \Omega = E\{1/\lambda_t (\partial \lambda_t/\partial \theta)(\partial \lambda_t/\partial \theta)^T\} \). To see this, we only need to verify assumptions (L1) and (L2). Note that by Fokianos et al. (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 \( h \geq 1 \), where \( \mu = EY_1 = \delta/(1-\alpha-\beta) \) and \( C(\theta) \) is a positive constant dependent on \( \theta \). Hence by monotone convergence theorem, we have

\[
E[Y_1 \log \{ \delta_U/\epsilon + \sum_{k=1}^{\infty} (1 - \epsilon)^k Y_{t-k} \}] \leq E[Y_1 \{ \delta_U/\epsilon + \sum_{k=1}^{\infty} (1 - \epsilon)^k Y_{t-k} \}] = \frac{\delta_U}{\epsilon} EY_1 + \sum_{k=1}^{\infty} (1 - \epsilon)^k EY_1 Y_{t-k} = \mu \frac{\delta_U}{\epsilon} + \sum_{k=1}^{\infty} (1 - \epsilon)^k \{ \gamma_Y(k) + \mu^2 \} < \infty.
\]

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

We now give the proof to Proposition 4 by applying the iterated random function approach.

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

\[
\begin{align*}
  f_u : [\lambda^*, \infty) \times [\lambda^*, \infty) \rightarrow [\lambda^*, \infty) \times [\lambda^*, \infty) \\
  \mathbf{x} = (\lambda_1, \lambda_2) \rightarrow (\lambda_2 + \alpha_1 \lambda_2 + \alpha_2 \lambda_1 + \beta F_{\lambda_2}^{-1}(u)).
\end{align*}
\]

For any \( \mathbf{x} = (x_1, x_2), \mathbf{y} = (y_1, y_2) \) in the state space \( S = [\lambda^*, \infty) \times [\lambda^*, \infty) \), define metric \( \rho \) as

\[
\rho(\mathbf{x}, \mathbf{y}) = w_1 |x_1 - y_1| + w_2 |x_2 - y_2|,
\]

where \( w_i > 0, i = 1, 2 \) and \( w_1, w_2 \) are to be decided. Let \( \mathbf{x}_1 = (\lambda_1^0, \lambda_2^0) := (\lambda^*, \lambda^*) \), then for any \( \mathbf{x} = (\lambda_1, \lambda_2) \) we have

\[
E\rho(\mathbf{X}_1(\mathbf{x}), \mathbf{X}_1(\mathbf{x}_1)) = \int_0^1 \rho(f_u(\mathbf{x}), f_u(\mathbf{x}_1)) du
= a_2 w_2 |\lambda_1 - \lambda_1^0| + \{w_1 + w_2(a_1 + b)|\lambda_2 - \lambda_2^0|,
\]

where the last equation 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

\[
E\rho(\mathbf{X}_1(\mathbf{x}), \mathbf{X}_1(\mathbf{x}_1)) \leq r\{w_1 |\lambda_1 - \lambda_1^0| + w_2 |\lambda_2 - \lambda_2^0|\}.
\]

This can be obtained if the equation \( r^2 - (a_1 + b)r - a_2 = 0 \) yields a root \( r_+ = \frac{a_1 + b + \sqrt{(a_1 + b)^2 + 4a_2}}{2} < 1 \). It can be shown that under \( \alpha_1 + \alpha_2 + \beta < 1 \) the root \( r_+ \in (0, 1) \). Note that the choice of \((w_1, w_2)\) is not unique.

2. \( q > 1 \): without loss of generality we consider the INGARCH(2,2) model. Define a Markov chain \( \mathbf{X}_t = (Y_t, \lambda_t, \lambda_{t+1}) \), then the chain can be obtained by defining the iterated random functions \( f_u : \mathbb{Z}_0 \times [\lambda^*, \infty) \times [\lambda^*, \infty) \rightarrow \mathbb{Z}_0 \times [\lambda^*, \infty) \times [\lambda^*, \infty) \) as \( f(\mathbf{x}) = f(n, \lambda_1, \lambda_2) = (F_{\lambda_2}^{-1}(u), \lambda_2 + \alpha_1 \lambda_2 + \alpha_2 \lambda_1 + \beta F_{\lambda_2}^{-1}(u) + \beta_2 n) \), where \( \lambda^* = \delta/(1 - \alpha_1 - \alpha_2) \) and \( u \in (0, 1) \). Note that we cannot define \( \mathbf{X}_t \) in the same way as in the first case, since otherwise it contradicts the independence assumption of \( \{u_t\} \) sequence. Define the metric \( \rho \) 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 \( w_i > 0, i = 1, 2, 3 \). Take \( \mathbf{x}_1 = (n_0, \lambda_1^0, \lambda_2^0) := (0, \lambda^*, \lambda^*) \), then for any \( \mathbf{x} = (n, \lambda_1, \lambda_2) \), we have

\[
E\rho(\mathbf{X}_1(\mathbf{x}), \mathbf{X}_1(\mathbf{x}_1)) = \int_0^1 |f_u(\mathbf{x}) - f_u(\mathbf{x}_1)| du
= \beta_{2} w_3 |n - n^0| + w_3 \alpha_2 |\lambda_1 - \lambda_1^0|
+ \{w_1 + w_2 + (\alpha_1 + \beta_1) w_3 |\lambda_2 - \lambda_2^0|.
\]

Similarly to 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.

\[\square\]
B.4. More on the NB-INGARCH

As defined in (4.7), the NB-INGARCH(1,1) model is given by

\[ Y_t | F_{t-1} \sim \text{NB}(r,p_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1}, \]

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, and has a unique stationary distribution \( \pi \); 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 assumption (L0), the maximum likelihood estimator \( \hat{\theta}_n \) is strongly consistent and asymptotically normal with mean \( \theta_0 \) and covariance matrix \( \Omega^{-1}/n \), where \( \Omega = \mathbb{E}[r/X_t/(X_t + r)(\partial X_t/\partial \theta)(\partial X_t/\partial \theta)^T] \). Verification of assumptions (L1) and (L2) is sufficient to demonstrate the result. Since \( B^{-1}(x) = \log \{x/(x+r)\} < 0 \), so assumption (L1) holds according to Remark 1. Note that \( A'(\eta) = re^{\eta}/(1-e^{\eta})^2 \) is increasing, so assumption (L2) holds provided \( \gamma_Y(0) < \infty \) 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

\[
\text{Var}(Y_0) = \mathbb{E}[\text{Var}(Y_0|X_0)] + \mathbb{E}[\text{Var}(Y_0|X_0)] = \mathbb{E}[r(1-p)/p_0^2] + \text{Var}(X_0) = \mu + 1/r \mathbb{E}X_0^2 + \text{Var}(X_0),
\]

and \( \text{Cov}(X_1, Y_1) = \mathbb{E}Y_1X_1 - \mu_1^2 = \mathbb{E}X_1^2 - \mu_2^2 = \text{Var}(X_1) \), 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 \( S \), \( \mathbb{E}[y_0 - f_u(y_0)] = \int_0^1 [y_0 - g(y_0, F_{y_0}^{-1}(u))]du \leq y_0 + g(0,0) + ay_0 + b \int_0^1 F_{y_0}^{-1}(u)du \leq y_0 + g(0,0) + (1 + a + b)y_0 < \infty \). Next for a fixed \( x_0 \in S \), there exists a unique \( y_0 \) such that \( x_0 = B(y_0) \) due to the strict monotonicity of \( B(\eta) \). For any \( x \geq x_0 \), there exists a unique \( \eta \geq y_0 \) such that \( x = B(\eta) \geq B(y_0) = x_0 \). Hence by the contraction condition (2.3), we have

\[
\mathbb{E}[X_1(x) - X_1(x_0)] = \int_0^1 |g(x, F_x^{-1}(u)) - g(x_0, F_{x_0}^{-1}(u))|du \leq a|x - x_0| + b\int_0^1 |F_x^{-1}(u) - F_{x_0}^{-1}(u)|du.
\]

(5.2)

It follows from \( x \geq x_0 \) and Proposition 6 that for any \( u \in (0,1) \), \( F_{x_0}^{-1}(u) \leq F_x^{-1}(u) \). Therefore

\[
\mathbb{E}[X_1(x) - X_1(x_0)] \leq a|x - x_0| + b\int_0^1 F_x^{-1}(u)du - \int_0^1 F_{x_0}^{-1}(u)du
\]

\[
= (a + b)(x - x_0).
\]

Similarly for \( x < x_0 \), we have \( \mathbb{E}[X_1(x) - X_1(x_0)] \leq (a + b)(x_0 - x) \). So for any \( x \in S \), we have \( \mathbb{E}[X_1(x) - X_1(x_0)] \leq (a + b)|x - x_0| \). Now suppose \( \mathbb{E}[X_n(x) - X_n(x_0)] \leq (a + b)^n|x - x_0| \), then

\[
\mathbb{E}[X_{n+1}(x) - X_{n+1}(x_0)] = \mathbb{E}[\mathbb{E} [X_{n+1}(X_n(x)) - X_{n+1}(X_n(x_0))]|U_1, \ldots, U_n] \leq E((a + b)|X_n(x) - X_n(x_0)|) \leq (a + b)^{n+1}|x - x_0|.
\]
By induction, \( \{X_t\} \) is geometric moment contracting and as a result, \( \pi \) is its unique stationary distribution.

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

\[
E(X_t|X_1) \leq \frac{1 - (a + b)^{t-1}}{1 - (a + b)} g(0, 0) + (a + b)^{t-1} X_1.
\]

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

\[
E_\pi X_1 \leq \liminf_{t \to \infty} E(X_t|X_1 = 0) \leq \frac{g(0, 0)}{1 - (a + b)} < \infty.
\]

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

### C.2. Proof of Proposition 2

Define a sequence of functions \( \{g_k, k \geq 1\} \) in a way such that \( g_1 = g \), and for \( k \geq 2 \), \( g_k(x, y_1, \ldots, y_k) = g_{k-1}(g(x, y_k), y_1, \ldots, y_{k-1}) \). 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 the contraction condition (2.3), we have \( E|X_t - g_1(0, Y_{t-1})| = E|g_1(X_{t-1}, Y_{t-1}) - g_1(0, Y_{t-1})| \leq aEX_{t-1} \). By induction, it follows that for any \( k \geq 1 \),

\[
E|X_t - g_k(0, Y_{t-1}, \ldots, Y_{t-k})| \leq a^kEX_{t-k}.
\]

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

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

\[
\beta(n) = E\{\sup_{A \in F_{-\infty,\infty}^Y} |P(A|F_{-\infty,0}^Y) - P(A)|\},
\]

where \( F_{-\infty,0}^Y = \sigma\{X_1, Y_0, Y_{-1}, \ldots\} \) according to (a). Because the distribution of \( (X_n, Y_{n+1}, \ldots) \) 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 \( n \geq 1 \), the coefficients of absolute regularity become

\[
\beta(n) = E\{\sup_{A \in F_{-\infty,\infty}^X} |P(A|\sigma\{X_1, Y_0, Y_{-1}, \ldots\}) - P(A)|\}
\]

\[
= E\{\sup_{A \in F_{-\infty,\infty}^X} |P(A|X_1) - P(A)|\}.
\]

(5.3)

Let \( \mathcal{B}^\infty \) be the \( \sigma \)-field in \( \mathbb{R}^\infty \) generated by the cylinder sets, then we can rewrite the coefficients of absolute regularity as

\[
\beta(n) = E\{\sup_{A \in \mathcal{B}^\infty} |P((X_n, Y_{n+1}, \ldots) \in A|X_1) - P((Y_n, Y_{n+1}, \ldots) \in A)|\}.
\]

(5.4)
We will provide an upper bound for (5.4) by coupling 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. Assume that both chains start from the stationary distribution, that is, \( X'_1 \sim \pi, X''_1 \sim \pi \) and that \( X'_1 \) is independent of \( X''_1 \). Let \( \{ U_k, k \in \mathbb{Z} \} \) be an iid sequence of uniform \((0, 1)\) random variables, and construct the chains as follows:

\[
X'_n = g(X'_{n-1}, F^{-1}_{X'_{n-1}}(U_{n-1})), \quad Y'_n = F^{-1}_{X'_n}(U_n), \\
X''_n = g(X''_{n-1}, F^{-1}_{X''_{n-1}}(U_{n-1})), \quad Y''_n = F^{-1}_{X''_n}(U_n).
\]

Since \( X'_1 \) and \( X''_1 \) are independent, so for any \( A \in B^\infty \),

\[
P((Y''_{n+1}, Y''_{n+1}, \ldots) \in A|X'_1) = P((Y'_n, Y_{n+1}, \ldots) \in A).
\]

Hence we have

\[
\left| P((Y_n, Y_{n+1}, \ldots) \in A|X_1 = x) - P((Y_n, Y_{n+1}, \ldots) \in A) \right| \\
= P((Y'_n, Y''_{n+1}, \ldots) \in A|X'_1 = x) - P((Y''_{n+1}, Y''_{n+1}, \ldots) \in A|X'_1 = x) \\
\leq P(Y'_n \neq Y''_n) \leq \sum_{k=0}^{\infty} P(Y'_{n+k} \neq Y''_{n+k}).
\]

Therefore the coefficients of absolute regularity are bounded by

\[
\beta(n) \leq P((Y'_n, Y'_{n+1}, \ldots) \neq (Y''_{n}, Y''_{n+1}, \ldots)) \leq \sum_{k=0}^{\infty} P(Y'_{n+k} \neq Y''_{n+k}).
\]

Observe that 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 \( E|X'_n - X''_n| \leq (a + b)^n \) for all \( n \). Then

\[
P(Y'_n \neq Y''_n) = E\{ P(Y'_n \neq Y''_n|X'_n, X''_n) \} = E\{ P(|Y'_n - Y''_n| \geq 1|X'_n, X''_n) \} \\
\leq E\{ E|Y'_n - Y''_n| |X'_n, X''_n) \} = E|X'_n - X''_n| \leq (a + b)^n.
\]

Hence according to (5.6), the coefficients of absolute regularity satisfy \( \beta(n) \leq \sum_{k=0}^{\infty} (a + b)^{n+k} = (a + b)^n/(1 - (a + b)) \). Recall the well-known fact that \( \beta \)-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 the 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. Note that by assumption (A1), \( X_t(\theta) \geq x^*_\theta \in \mathcal{R}(B) \), which implies \( \eta_1(\theta) \geq B^{-1}(x^*_\theta) \). So it follows from assumptions (A2) and (A4) that for any \( \theta \in \Theta \),

\[
E l_t(\theta) = E\{ Y_t B^{-1}(X_t(\theta)) - A(B^{-1}(X_t(\theta))) \} \\
\leq E\{ Y_t \sup_{\theta \in \Theta} B^{-1}(X_t(\theta)) \} - A((B^{-1}(x^*_\theta)) < \infty.
\]

This implies \( E l^*_t(\theta) < \infty \). Denote \( M_n(\theta) = \sum_{t=1}^{n} l_t(\theta)/n \), then \( M_n(\theta) \xrightarrow{a.s.} M(\theta) = E\{ Y_1 \eta_1(\theta) - A(\eta_1(\theta)) \} \) according to the extended mean ergodic theorem (see Billingsley (1995) pp. 284 and
In order to prove the identifiability, we need to show that $\theta_0$ is the unique maximizer of $M(\theta)$, that is, for any $\theta \in \Theta \setminus \{\theta_0\}$, $M(\theta) - M(\theta_0) < 0$. First it follows from assumption (A5) that for any $\theta \neq \theta_0$ and all $t$, $P_{\theta_0}(G_t(\theta, \theta_0)) > 0$, where $G_t(\theta, \theta_0) = \{X_t(\theta) \neq X_t(\theta_0)\}$. Let $G = G_t(\theta, \theta_0)$, then we have

$$
M(\theta) - M(\theta_0) = E_Y \left\{ B^{-1}(X_t(\theta)) - B^{-1}(X_t(\theta_0)) \right\} - \{ A(B^{-1}(X_t(\theta))) - A(B^{-1}(X_t(\theta_0))) \} = E_Y \left\{ X_t(\theta) - X_t(\theta_0) \right\} - \{ A(B^{-1}(X_t(\theta))) - A(B^{-1}(X_t(\theta_0))) \} = \int_G X_t(\theta) - X_t(\theta_0) - A(B^{-1}(X_t(\theta))) - A(B^{-1}(X_t(\theta_0))) dP_{\theta_0}.
$$

On the set $G$, there exists $c \in \mathbb{R}$ between $B^{-1}(X_t(\theta))$ and $B^{-1}(X_t(\theta_0))$ 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. It follows from $A''(\eta) > 0$ that $A(\eta)$ 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

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

Since $B(\eta)$ is strictly increasing, so $(X_t(\theta_0) - \xi)\{ B^{-1}(X_t(\theta)) - B^{-1}(X_t(\theta_0)) \} < 0$ in either of the two cases: $X_t(\theta) < X_t(\theta_0)$ and $X_t(\theta) > X_t(\theta_0)$. Hence $M(\theta) - M(\theta_0) < 0$, for any $\theta \neq \theta_0$, which establishes the identifiability. To show the consistency, first note that by assumption (A4), we have

$$
E \sup_{\theta \in \Theta} l_t(\theta) = E \{ Y_t \sup_{\theta \in \Theta} B^{-1}(X_t(\theta)) - \inf_{\theta \in \Theta} A(B^{-1}(X_t(\theta))) \} \leq E \{ Y_t \sup_{\theta \in \Theta} B^{-1}(X_t(\theta)) \} - A(B^{-1}(x^*)) < \infty.
$$

The function $f_\theta$ in Lemma 1 can be defined as

$$
f_\theta(y) = y_1 B^{-1}(y_\infty(y_0, y_{-1}, \ldots)) - A(B^{-1}(g_\infty(y_0, y_{-1}, \ldots))),
$$

where $y = (y_1, y_0, y_{-1}, \ldots)$. Hence it follows from assumption (A2) and Lemma 1 that $M(\theta)$ is upper-semicontinuous and for any compact subset $K \subset \Theta$, $\limsup_{n \to \infty} \sup_{\theta \in K} M_n(\theta) = \sup_{\theta \in K} M(\theta)$. Take $U_0$ as a local base of $\theta_0$ and let $U \in U_0$ be a neighborhood of $\theta_0$, then Lemma 1 can be applied to $\Theta \setminus U$. Because u.s.c function attains its maximum on compact sets and $M(\theta) < M(\theta_0)$ for any $\theta \neq \theta_0$, we have

$$
\limsup_{n \to \infty} \sup_{\theta \in \Theta \setminus U} M_n(\theta) \leq \sup_{\theta \in \Theta \setminus U} M(\theta) < M(\theta_0), \quad P_{\theta_0} \text{-a.s.} \quad (5.7)
$$

Notice that for any $\hat{\theta} \notin U$, $M_n(\hat{\theta}) \leq \sup_{\theta \in \Theta \setminus U} M_n(\theta)$. Let $\omega \in \Omega$ such that (5.7) holds and $M(\theta_0) = \lim_{n \to \infty} M_n(\theta_0)$. For such $\omega$, suppose $\hat{\theta}_n \notin U$ infinitely often, say, along a sequence denoted by $N$, then

$$
\liminf_{n \to \infty} M_n(\hat{\theta}_n) \leq \liminf_{n \to \infty, n \in N} M_n(\hat{\theta}_n) \leq \limsup_{n \to \infty, n \in N} M_n(\hat{\theta}_n) \leq \limsup_{n \to \infty, n \in N} \sup_{\theta \notin U} M_n(\theta) \leq \limsup_{n \to \infty} \sup_{\theta \notin U} M_n(\theta).
$$
However, according to (5.7), we have

\[
\limsup_{n \to \infty} \sup_{\theta \in \Theta \cup U} M_n(\theta) \leq \sup_{\theta \in \Theta \setminus U} M(\theta) = \lim_{n \to \infty} M_n(\theta_0) \leq \liminf_{n \to \infty} M_n(\hat{\theta}_n),
\]

which contradicts (5.8). 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 \( \theta_0 \) almost surely.

### C.4. Proof of Theorem 2

We define a linearized form of \( \eta_t(\theta) \) as \( \eta_t^\dagger(\theta) := \eta_t(\theta_0) + (\theta - \theta_0)^T \eta_t \), and the corresponding linearized log-likelihood function of \( l(\theta) \) as

\[
l_t^\dagger(\theta) := \sum_{t=1}^{n} \eta_t^\dagger(\theta) Y_t - \sum_{t=1}^{n} A(\eta_t^\dagger(\theta)).
\]

Let \( u = \sqrt{n}(\theta - \theta_0) \), then define

\[
R_n^\dagger(u) = l_t^\dagger(\theta_0) - l_t^\dagger(\theta_0 + un^{-1/2}) = \sum_{t=1}^{n} Y_t \eta_t - \sum_{t=1}^{n} A(\eta_t) - \sum_{t=1}^{n} (\eta_t + u^T n^{-1/2} \eta_t) Y_t + \sum_{t=1}^{n} A(\eta_t + u^T n^{-1/2} \eta_t)
\]

\[
= -u^T n^{-1/2} \sum_{t=1}^{n} Y_t \eta_t + \sum_{t=1}^{n} \{A(\eta_t + u^T n^{-1/2} \eta_t) - A(\eta_t)\}
\]

\[
= -u^T n^{-1/2} \sum_{t=1}^{n} \{Y_t - B(\eta_t)\} \eta_t + \sum_{t=1}^{n} \{A(\eta_t + u^T n^{-1/2} \eta_t) - A(\eta_t) - u^T n^{-1/2} B(\eta_t) \eta_t\}.
\]

(5.9)

Let \( s_t = n^{-1/2}\{Y_t - B(\eta_t)\} \eta_t \), then \( E(s_t|\mathcal{F}_{t-1}) = n^{-1/2} E\{Y_t - B(\eta_t)\} \eta_t|\mathcal{F}_{t-1} = 0 \), so \( \{s_t, t \geq 1\} \) is a martingale difference sequence. Note that

\[
\sum_{t=1}^{n} E(s_t s_t^T | \mathcal{F}_{t-1}) = \frac{1}{n} \sum_{t=1}^{n} E\{Y_t - B(\eta_t)\}^2 \eta_t \eta_t^T | \mathcal{F}_{t-1}
\]

\[
= \frac{1}{n} \sum_{t=1}^{n} B'(\eta_t) \eta_t \eta_t^T,
\]

which converges almost surely to \( \Omega \) by the mean ergodic theorem and assumption (A7). Moreover, for any \( \epsilon > 0 \),

\[
\sum_{t=1}^{n} E\{s_t s_t^T 1_{||s_t|| \geq \epsilon}|\mathcal{F}_{t-1}\}
\]

\[
= \frac{1}{n} \sum_{t=1}^{n} \eta_t \eta_t^T E\{Y_t - B(\eta_t)\}^2 1_{||Y_t - B(\eta_t)\eta_t|| \geq \epsilon \sqrt{n}} | \mathcal{F}_{t-1}
\]

\[
\leq \frac{1}{n} \sum_{t=1}^{n} \eta_t \eta_t^T E\{Y_t - B(\eta_t)\}^2 1_{||Y_t - B(\eta_t)\eta_t|| \geq M} | \mathcal{F}_{t-1}
\]

\[
\rightarrow E\{Y_t - B(\eta_t)\}^2 \eta_t \eta_t^T 1_{||Y_t - B(\eta_t)\eta_t|| \geq M} \quad \text{as} \ n \to \infty
\]

\[
\rightarrow 0 \quad \text{as} \ M \to 0.
\]
Then it follows from the central limit theorem for martingale difference sequences that
\[
\sum_{t=1}^{n} s_t \xrightarrow{c} V \sim N(0, \Omega), \quad \text{as} \quad n \to \infty,
\]
where \( \Omega \) is evaluated at \( \theta_0 \). The other term in (5.9) by Taylor expansion is
\[
\frac{1}{2n} \sum_{t=1}^{n} u^T \{ B'(\eta_t) \dot{\eta}_t \dot{\eta}_t^T \} u + \mathcal{O}_p(n^{-3/2}) \sum_{t=1}^{n} B''(\eta_t) (u^T \dot{\eta}_t)^3,
\]
which is of the order of \( u^T \Omega u/2 + o_P(1) \). Hence \( R_n^1(u) \xrightarrow{c} -u^T V + \frac{1}{2} u^T \Omega u \), where \( V \sim N(0, \Omega) \).
It then follows that \( \arg\min_u R_n^1(u) \xrightarrow{c} \arg\min_u \{ -u^T V + \frac{1}{2} u^T \Omega u \} = \Omega^{-1} V \sim N(0, \Omega^{-1}) \).

For the rest of the proof, we show that the difference between \( R_n(u) := l(\theta_0) - l(\theta_0 + un^{-1/2}) \) and \( R_n^1(u) \) is negligible as \( n \) grows large. By writing \( \theta = \theta_0 + un^{-1/2} \), the difference becomes
\[
R_n^1(u) - R_n(u) = \sum_{t=1}^{n} \{ Y_t - B(\eta_t) \} \{ \eta_t(\theta) - \eta_t - u^T n^{-1/2} \dot{\eta}_t \}
- \sum_{t=1}^{n} [A(\eta_t(\theta)) - A(\eta_t + u^T n^{-1/2} \dot{\eta}_t)]
- B(\eta_t) \{ \eta_t(\theta) - \eta_t - u^T n^{-1/2} \dot{\eta}_t \}. \tag{5.10}
\]
By Taylor expansion, the first term in (5.10) is \( 1/(2n) \sum_{t=1}^{n} \{ Y_t - B(\eta_t) \} u^T \dot{\eta}_t (\dot{\theta}_t^*) u = 1/(2n) u^T \sum_{t=1}^{n} \{ Y_t - B(\eta_t) \} \dot{\eta}_t (\dot{\theta}_t^*) u \), where \( \dot{\theta}_t^* \) lies between \( \theta \) and \( \theta_0 \), and \( \dot{\eta}_t = \partial^2 \eta_t / \partial \theta \partial \theta^T \). Since
\[
\frac{1}{n} \sum_{t=1}^{n} \{ Y_t - B(\eta_t) \} \dot{\eta}_t \overset{a.s.}{\longrightarrow} \mathbb{E}[\{ Y_t - B(\eta_t) \} \dot{\eta}_t]
= \mathbb{E}[\dot{\eta}_t \mathbb{E}[Y_t - B(\eta_t) | F_{t-1}] | \dot{\eta}_t] = 0,
\]
and \( 1/n \sum_{t=1}^{n} \{ Y_t - B(\eta_t) \} \dot{\eta}_t (\dot{\theta}_t^*) \overset{a.s.}{\longrightarrow} 0 \) under the smoothness assumption, so the first term in (5.10) converges to 0 uniformly on \([-K, K]\) for any \( K > 0 \). We now apply Taylor expansion to each component in the second term of (5.10),
\[
A(\eta_t(\theta)) = A(\eta_t) + u^T n^{-1/2} B(\eta_t) \dot{\eta}_t + \frac{1}{2n} u^T \{ B(\eta_t(\dot{\theta}_t^*)) \dot{\eta}_t (\dot{\theta}_t^*) + B'(\dot{\theta}_t^*) \dot{\eta}_t (\dot{\theta}_t^*) \dot{\eta}_t (\dot{\theta}_t^*)^T \} u,
\]
\[
A(\eta_t + u^T n^{-1/2} \dot{\eta}_t) = A(\eta_t) + B(\eta_t) u^T n^{-1/2} \dot{\eta}_t + \frac{1}{2n} u^T B'(c) \dot{\eta}_t \dot{\eta}_t^T u,
\]
\[
\eta_t(\theta) = \eta_t(\theta_0 + un^{-1/2}) = \eta_t + \dot{\eta}_t u^T n^{-1/2} + \frac{1}{2n} u^T \dot{\eta}_t (\dot{\theta}_2^*) u,
\]
where \( 0 \leq c \leq u^T n^{-1/2} \dot{\eta}_t \), \( \dot{\theta}_1^* \) and \( \dot{\theta}_2^* \) both lie between \( \theta_0 \) and \( \theta \). Therefore the second term in (5.10)
becomes
\[
\sum_{t=1}^{n} \left[ A(\eta_t(\theta)) - A(\eta_t + u^T n^{-1/2} \hat{\eta}_t) - B(\eta_t)\{\eta_t(\theta) - \eta_t - u^T n^{-1/2} \hat{\eta}_t\} \right]
\]
\[
= \sum_{t=1}^{n} \left[ A(\eta_t) + u^T n^{-1/2} B(\eta_t) \hat{\eta}_t + \frac{1}{2n} u^T \left\{ B(\eta_t(\theta_1^*)) \hat{\eta}_t(\theta_1^*) + B'(\theta_1^*) \hat{\eta}_t(\theta_1^*) \hat{\eta}_t(\theta_1^*)^T \right\} u - A(\eta_t) - B(\eta_t) u^T n^{-1/2} \hat{\eta}_t - \frac{1}{2n} u^T B'(c) \hat{\eta}_t \hat{\eta}_t^T u \right]
\]
\[
= \frac{1}{2n} u^T \sum_{t=1}^{n} \left\{ [B(\eta_t(\theta_1^*)) \hat{\eta}_t(\theta_1^*) - B(\eta_t) \hat{\eta}_t(\theta_2^*)] + \{B'(\theta_1^*) \hat{\eta}_t(\theta_1^*) \hat{\eta}_t(\theta_1^*)^T - B'(c) \hat{\eta}_t \hat{\eta}_t^T \} u \right\},
\]
which converges to 0 on a compact set of \( u \) under smoothness assumptions. So (5.10) converges to 0 as \( n \to \infty \), which implies that \( \arg\min_u R_n(u) \) and \( \arg\min_u R_n^u(u) \) have the same asymptotic distribution, i.e.,
\[
\arg\min_u R_n(u) \xrightarrow{\mathcal{L}} \Omega^{-1} V \sim N(0, \Omega^{-1}).
\]
Note that \( \arg\min_u R_n(u) = \arg\max_u l(\theta_0 + un^{-1/2}) = \sqrt{n}(\hat{\theta}_n - \theta_0) \), where \( \hat{\theta}_n \) is the conditional maximum likelihood estimator. Hence
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega^{-1}), \quad \text{as} \quad n \to \infty.
\]

C.5. Proof of Theorem 3

According to Theorems 1 and 2, it is sufficient to establish the identifiability of the model, that is, we need to verify assumption (A5). Suppose for some \( t \in \mathbb{Z} \), \( X_t(\theta) = X_t(\theta_0) \), \( P_{\theta_0}\)-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 (5.1) that
\[
(\beta - \beta_0) Y_{t-1} = \delta_0 - \delta + \alpha_0 \sum_{k=0}^{\infty} \alpha_0^k Y_{t-k-2} - \alpha(\frac{\delta}{1 - \alpha} + \beta \sum_{k=0}^{\infty} \alpha^k Y_{t-k-2}).
\]
If \( \beta \neq \beta_0 \), then \( Y_{t-1} \in \text{span}\{Y_{t-2}, Y_{t-3}, \ldots\} \) which contradicts the fact that \( \text{Var}(Y_{t-1}|\mathcal{F}_{t-2}) > 0 \). So \( \beta \) must be the same as \( \beta_0 \). 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 only need to 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 \( \theta = \theta_0 \) which verifies the identifiability of the model.
References


