Some Models for Time Series of Counts

Heng Liu

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

ABSTRACT

Some Models for Time Series of Counts

Heng Liu

This thesis focuses on developing nonlinear time series models and establishing relevant theory with a view towards applications in which the responses are integer valued. The discreteness of the observations, which is not appropriate with classical time series models, requires novel modeling strategies. The majority of the existing models for time series of counts assume that the observations follow a Poisson distribution conditional on an accompanying intensity process that drives the serial dynamics of the model. According to whether the evolution of the intensity process depends on the observations or solely on an external process, the models are classified into parameter-driven and observation-driven. Compared to the former one, an observation-driven model often allows for easier and more straightforward estimation of the model parameters. On the other hand, the stability properties of the process, such as the existence and uniqueness of a stationary and ergodic solution that are required for deriving asymptotic theory of the parameter estimates, can be quite

complicated to establish, as compared to parameter-driven models.

In this thesis, we first propose a broad class of observation-driven models that is based upon a one-parameter exponential family of distributions and incorporates nonlinear dynamics. The establishment of stability properties of these processes, which is at the heart of this thesis, is addressed by employing theory from iterated random functions and coupling techniques. Using this theory, we are also able to obtain the asymptotic behavior of maximum likelihood estimates of the parameters.

Extensions of the base model in several directions are considered. Inspired by the idea of a self-excited threshold ARMA process, a threshold Poisson autoregression is proposed. It introduces a two-regime structure in the intensity process and essentially allows for modeling negatively correlated observations. E-chain, a non-standard Markov chain technique and Lyapunov's method are utilized to show the stationarity and a law of large numbers for this process. In addition, the model has been adapted to incorporate covariates, an important problem of practical and primary interest.

The base model is also extended to consider the case of multivariate time series of counts. Given a suitable definition of a multivariate Poisson distribution, a multivariate Poisson autoregression process is described and its properties studied.

Several simulation studies are presented to illustrate the inference theory. The proposed models are also applied to several real data sets, including the number of transactions of the Ericsson stock, the return times of Goldman Sachs Group stock prices, the number of road crashes in Schiphol, the frequencies of occurrences of gold particles, the incidences of polio in the US and the number of presentations of asthma in an Australian hospital. An array of graphical and quantitative diagnostic tools,

which is specifically designed for the evaluation of goodness of fit for time series of counts models, is described and illustrated with these data sets.

Contents

List of Tables

List of Figures

Acknowledgments

I owe my deeply-felt thanks to my Ph.D advisor, Professor Richard A. Davis, for his constant encouragement, support and guidance throughout my graduate study and my thesis-writing period. Prof. Davis has been not only a great academic advisor, but a great mentor and company in my life. He has taught me so much that I can still benefit greatly from for the years to come.

I would also like to thank Professors Liam Paninski, Jingchen Liu, Karl Sigman and Srikesh Arunajadai for agreeing to serve on my defense committee.

I am grateful to all my friends at the Department of Statistics at Columbia University for their constant encouragement and support throughout the past four years. I have had the opportunity to learn from many professors in the department and got inspiration from my fellow Ph.D classmates.

Last but most importantly, I want to express my deepest thanks to my parents and grandparents, for their tolerance, understanding and love.

To my parents, grandparents and those who educate me

Chapter 1

Introduction

1.1 Background

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. The majority of these models assume that the observations follow a Poisson distribution conditional on an accompanying intensity process that drives the dynamics of the model, e.g., Zeger (1988), Campbell (1994), Streett (2000), Davis *et al.* (2003), Fokianos *et al.* (2009), Neumann (2011) 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. Most of the popular models used these days can be categorized into one of these two classes. This thesis mainly focuses on developing theory and inference for a variety of observation-driven models.

For a parameter-driven model, one usually needs to resort to simulation-based numerical methods to obtain parameter estimates, which is more than often rather

computationally intensive. An observation-driven model, on the other hand, enjoys a considerably easier and more straightforward estimation procedure; however, it is difficult to establish stability properties of the model, including stationarity and mixing conditions. In the next two sections, some well-known model examples are provided to shed some light on the characteristics of both categories of the models. For illustration purposes, only the models based on a Poisson distribution are provided in this chapter, while some of them can be generalized to be based upon other discrete distributions and will be discussed in details in this thesis.

1.2 Parameter-Driven Models

Let Y_t and x_t denote the observation and the explanatory regression vector at time t, respectively. Then a parameter-driven model assumes that Y_t follows a Poisson distribution given the intensity

$$\mu_t := \exp\{x_t^T\beta + \alpha_t\}, \tag{1.2.1}$$

where $\{\alpha_t\}$ is a stationary Gaussian process and β is the vector of regression coefficients. The variants of the model come from different types of structure imposed on $\{\alpha_t\}$. A simple but illustrative example is that $\{\alpha_t\}$ itself is an AR(1) process, i.e.,

$$\alpha_t + \sigma^2/2 = \phi(\alpha_{t-1} + \sigma^2/2) + \epsilon_t, \tag{1.2.2}$$

where $\{\epsilon_t\} \sim \text{IID}(0, \sigma^2(1-\phi^2))$. In the special case that $x_t^T\beta = \beta$, i.e., $\mu_t = \exp\{\beta + \alpha_t\}$, the stability properties are easy to derive, since μ_t can be easily seen as a function of $\epsilon_t, \epsilon_{t-1}, \ldots$. In addition, the regression parameters in model (1.2.1) are interpretable on the log mean scale. To see this, note that it follows from (1.2.2) that $\text{E}[\exp\{\alpha_t\}] = 1$, hence

$$\text{E}Y_t = \exp\{x_t^T\beta\}\text{E}\exp\{\alpha_t\} = \exp\{x_t^T\beta\}.$$

However, the estimation of the parameters proves to be difficult, which is due to the fact that the likelihood function involves an n-fold integral, where n is the sample size. Observe that the model specification (1.2.1) falls into the framework of a non-Gaussian state-space model, and there has been a vast volume of literature studying the associated estimation problem, see for example Chan and Ledolter (1995), Kitagawa (1996), Durbin and Koopman (2001) and Davis and Rodriguez-Yam (2005). More recently, Davis and Yao (2009) considered a pair-wise likelihood method to estimate the parameters and investigated the performance of such an approximation.

1.3 Observation-Driven Models

Unlike the parameter-driven model (1.2.1), the conditional mean μ_t in an observation-driven model depends explicitly on previous observations. The estimation becomes easier and more straightforward, since the likelihood function can be computed in a direct and recursive fashion. Nevertheless, some theoretical issues, including establishing stationarity and ergodicity of the processes, can be difficult to resolve.

Among all the observation-driven models, the generalized linear ARMA process (GLARMA), which was first proposed by Davis *et al.* (2003), has received extensive study in the literature. It is assumed that

$$Y_t|\mu_t \sim \text{Poisson}(\mu_t), \quad \text{where } \log \mu_t = x_t^T \beta + \alpha_t,$$

where $\alpha_t = \sum_{i=1}^{\infty} \psi_i e_{t-i}$, and $\{e_t\}$ is a sequence of martingale differences, which is defined as

$$e_t = (Y_t - \mu_t)/\mu_t^{\lambda},$$

where $\lambda \geq 0$ and usually takes values $0, 1/2$ or 1. Recently, Blasques *et al.* (2012) have considered a generalized autoregressive score model, which includes the GLARMA as

a special case under some constraints on the parameter space. However, only under very restrictive conditions on the parameters have the stability properties of the GLARMA process been established and relevant study is still ongoing.

Another well-known observation-driven model is the Poisson integer-valued generalized autoregressive conditional heteroscedastic process (INGARCH), also known as a Poisson autoregression (see e.g., Ferland *et al.* (2006), Fokianos *et al.* (2009) and Davis and Liu (2012)). It is assumed that the conditional mean μ_t of the observation Y_t is a linear combination of the previous conditional mean and the previous observation, i.e., $\mu_t = d + a\mu_{t-1} + bY_{t-1}$. To maintain the positivity of μ_t, all the parameters are restricted to be non-negative. In particular, it has been shown in the literature that the model is stationary and satisfies some mixing conditions if $a + b < 1$, see for example Neumann (2011), Davis and Liu (2012) and Doukhan *et al.* (2012). The model is known for its capability of capturing serial dependence and easy implementation of likelihood calculation. The details will be postponed to the next chapter, since it is actually a special case of the model that will be introduced there.

1.4 Main Results

This thesis mainly focuses on developing theory and inference for observation-driven models. We first propose a broad class of observation-driven models that is based upon a one-parameter exponential family of distributions and incorporates nonlinear dynamics, which introduces extra flexibility when fitting the model to real applications. The establishment of stability properties of the processes is addressed by employing theory from iterated random functions (see Appendix) and a special coupling technique under some constraints on the parameter space. This in turn helps develop the asymptotic theory of the parameter estimates. Examples of both linear

and nonlinear dynamic models are presented.

Extensions of the base model in several directions are considered in order to improve its practicality. Inspired by the idea of a self-excited threshold ARMA process proposed by Tong (1990), a self-excited threshold Poisson autoregression is proposed. It introduces a two-regime structure in the intensity process and essentially allows for modeling negatively correlated observations. E-chain, a non-standard Markov chain technique (see Definition 6.2.10 and Theorem 6.2.2) and Lyapunov's method are utilized to show the stationarity and a law of large numbers for this process. In addition, the model has been adapted to incorporate covariates, an important problem of practical and primary interest.

The base model is also extended to consider the case of multivariate time series of counts. Given a suitable definition of a multivariate (bivariate) Poisson distribution, a multivariate (bivariate) Poisson autoregression process is described and its properties studied.

Several simulation studies are presented to illustrate the inference theory developed in this thesis. The proposed models are also applied to several real data sets, including the number of transactions of the Ericsson stock, the return times of Goldman Sachs Group stock prices, the number of road crashes in Schiphol, the frequencies of occurrences of gold particles, the incidences of polio in the US and the number of presentations of asthma in an Australian hospital. An array of graphical and quantitative diagnostic tools, which is specifically designed for the evaluation of goodness of fit for time series of counts models, is described and illustrated with these data sets.

1.5 Organization of the Thesis

Chapter 2 develops theory and inference for a particular class of observation-driven models for time series of counts. Some linear and nonlinear dynamic models are considered and two data applications are presented in this chapter too. Extensions in a variety of directions of this class of models are discussed and investigated in Chapter 3, where several new data applications are provided. In Chapter 4, a bivariate Poisson integer-valued GARCH model is studied and comparison with other models is drawn based on a real data application. Chapter 5 summarizes some key Markov chain theory that is used throughout the thesis.

Chapter 2

A Class of Nonlinear Models

2.1 Introduction

This chapter focuses on the theory and inference for a particular class of observation-driven models. Many of the commonly used models, such as the Poisson integer-valued GARCH (INGARCH), are special cases of our model. The INGARCH model, also known as the Poisson autoregression, has already received considerable study in the literature, see for example, Ferland *et al.* (2006), Fokianos *et al.* (2009), Neumann (2011), Doukhan *et al.* (2012) and Davis and Liu (2012). For this model, it is assumed that the observations $\{Y_t\}$ given the intensity process $\{\lambda_t\}$ follow a Poisson distribution, where λ_t follows the GARCH-like recursions $\lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1}$. The model is named after the usual GARCH model (Bollerslev (1986)) since the Poisson mean coincides with its variance, and is known for its capability of capturing positive temporal dependence in the observations and it is relatively easy to fit via maximum likelihood. Fokianos *et al.* (2009) studied the model and established the asymptotic theory of the parameter estimates by introducing a small perturbation and Neumann

(2011) considered some contracting dynamics of λ_t and derived mixing condition of the count process. Davis and Liu (2012) generalized the conditional distribution of $\{Y_t\}$ to a one-parameter exponential family and took advantage of the theory from iterated random functions (Diaconis and Freedman (1999) and Wu and Shao (2004)) to establish stationarity and absolute regularity of the process, as well as the asymptotic distribution of the parameter estimates. Doukhan *et al.* (2012) showed similar results by utilizing the concept of τ-weak dependence. More recently, Blasques *et al.* (2012) considered a class of generalized autoregressive score processes which includes the INGARCH 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.

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 Appendix for details, 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 this chapter is as follows. Section 2.2 formulates the model and establishes stability properties. The maximum likelihood estimates of the parameters and the relevant asymptotic theory are derived in Section 2.3. Examples of both linear and nonlinear dynamic models are considered in Section 2.4. Numerical results, including a simulation study and two data applications are given in Section 2.5, 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 2.5.

2.2 Model Formulation and Stability Properties

2.2.1 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

$$p(y|\eta) = \exp\{\eta y - A(\eta)\}h(y), \quad y \geq 0, \tag{2.2.1}$$

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 $A(\eta)$ exists generally 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 the thesis, so $B(\eta) = \mathrm{E}Y > 0$, which implies that $A(\eta)$ is strictly increasing.

An important property of the one-parameter exponential family that is heavily used in our research is the stochastic monotonicity. A random variable X is said to be stochastically smaller than a random variable Y (written as $X \leq_{ST} \mathrm{Y}$) 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 2.2.1. *Suppose two random variables Y' and Y'' follow distributions belonging to the one-parameter exponential family (2.2.1) 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. Denote the probability density functions of Y' and Y'' as $p(y|\eta')$ and $p(y|\eta'')$ defined in (2.2.1), respectively. Then the log ratio of the two densities is

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

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., $Y' \leq_{lc} Y''$. Moreover, since $A(\eta)$ is increasing in η, so $\lim_{y \downarrow 0} l(y) = A(\eta'') - A(\eta') \geq 0$ 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'', i.e., $Y' \leq_{ST} Y''$. □

Denote F_x as the cumulative distribution function of $p(y|\eta)$ in (2.2.1) 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 2.2.2. *Suppose that U is a uniform $(0, 1)$ random variable, and define two random variables Y' and Y'' as*

$$Y' = F_{x'}^{-1}(U) \quad \textit{and} \quad Y'' = F_{x''}^{-1}(U),$$

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 (2.2.1) 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 2.2.1. It follows that $F^{-1}_{x'}(\theta) \leq F^{-1}_{x''}(\theta)$ 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 $\mathrm{E}|Y' - Y''| = |x' - x''|$. □

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.

2.2.2 Model Formulation

Set $\mathcal{F}_0 = \sigma\{\eta_1\}$, where η_1 is a natural parameter of (2.2.1) 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,

$$Y_t|\mathcal{F}_{t-1} \sim p(y|\eta_t), \quad X_t = g_\theta(X_{t-1}, Y_{t-1}), \tag{2.2.2}$$

for all $t \geq 1$, where $p(y|\eta_t)$ is defined in (2.2.1), $\mathcal{F}_t = \sigma\{\eta_1, Y_1, \ldots, Y_t\}$ and X_t is the conditional mean process, i.e., $X_t = B(\eta_t) = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$. 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, i.e., for any $x, x' \geq 0$, and $y, y' \in [0,\infty)$ or $\mathbb{N}_0$,

$$|g_\theta(x, y) - g_\theta(x', y')| \leq a|x - x'| + b|y - y'|, \tag{2.2.3}$$

where a and b are non-negative constants with $a + b < 1$. Note that (2.2.3) implies

$$g_\theta(x, y) \leq g_\theta(0, 0) + ax + by, \quad \text{for any} \quad x, y \geq 0. \tag{2.2.4}$$

We point out that model (2.2.2) with the function g_θ satisfying (2.2.3) includes the Poisson INGARCH model (see Example 2.4.1) and the exponential autoregressive model (2.4.13) 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 chapter is on the conditional mean process $\{X_t\}$, which can be easily seen as a time-homogeneous Markov chain. Note that the observation process $\{Y_t\}$ is not a Markov chain itself.

2.2.3 Stationarity and Mixing Conditions

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. The relevant definitions and theorems are introduced in Appendix. We will demonstrate that the conditional mean process $\{X_t\}$ specified in (2.2.2) can be embedded into the framework of iterated random function approach (IRF) and shown to be geometric moment contracting (GMC).

In this section and the next we use g to represent the function g_θ in (2.2.2) evaluated at the true parameter. For any $u \in (0, 1)$, the random function $f_u(x)$ is defined as

$$f_u(x) := g\big(x, F_x^{-1}(u)\big), \tag{2.2.5}$$

where F_x is the cumulative distribution function of $p(y|\eta)$ in (2.2.1) 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 (2.2.2) starting from $X_0 = x$ can be represented as the so-called forward process $X_t(x) = (f_{U_t} \circ f_{U_{t-1}} \circ \ldots \circ f_{U_1})(x)$ (see (6.3.1)). 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 2.2.3. *Assume model (2.2.2) and that the function g satisfies the contraction condition (2.2.3). Then*

1. *There exists a random variable Z_∞ such that, for all $x \in S$, $Z_n(x) \to Z_\infty$ almost surely. The limit Z_∞ does not depend on x and has distribution π, which is 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. In addition, $E_\pi X_1 < \infty$.*

3. *If $\{X_t, t \geq 1\}$ starts from π, i.e., $X_1 \sim \pi$, then $\{Y_t, t \geq 1\}$ is a stationary time series.*

Proof. According to Theorem 6.3.2, it suffices to verify Conditions 1 and 2 formulated as (6.3.3) and (6.3.4). For any y_0 in the state space S, $\mathrm{E}|y_0 - f_u(y_0)| = \int_0^1 |y_0 - g(y_0, F_{y_0}^{-1}(u))|du \leq y_0 + g(0, 0) + ay_0 + b\int_0^1 F_{y_0}^{-1}(u)du \leq g(0, 0) + (1 + a + b)y_0 < \infty$. 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 the contraction condition (2.2.3), we have

$$\begin{aligned} \mathrm{E}|X_1(x) - X_1(x_0)| &= \int_0^1 \big|g\big(x, F_x^{-1}(u)\big) - g\big(x_0, F_{x_0}^{-1}(u)\big)\big| du \\ &\leq a|x - x_0| + b\int_0^1 \big|F_x^{-1}(u) - F_{x_0}^{-1}(u)\big| du. \end{aligned} \tag{2.2.6}$$

It follows from $x \geq x_0$ and Proposition 2.2.1 that for any $u \in (0,1)$, $F_{x_0}^{-1}(u) \leq F_x^{-1}(u)$. Therefore

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

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

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

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

To show that $\mathrm{E}_\pi X_1 < \infty$, notice that by taking conditional expectation on both sides of (2.2.4), we have $\mathrm{E}(X_t|X_{t-1}) \leq g(0,0) + (a+b)X_{t-1}$. Inductively one can show that for any $t \geq 1$,

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

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

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

To prove (3), 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 $Y_t = F^{-1}_{X_t}(\xi_t)$. 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. $\square$

Proposition 2.2.3 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 $n \to \infty$. To further investigate the stability properties, including ergodicity and mixing conditions for model (2.2.2), we extend $\{(X_t, Y_t)\}$ to be indexed by all the integers since it is strictly stationary under the conditions of Proposition 2.2.3. The following proposition establishes ergodicity and absolute regularity when Y_t is discrete.

Proposition 2.2.4. *Assume model (2.2.2) where the support of Y_t is a subset of $\mathbb{N}_0 = \{0, 1, \ldots, \}$, and that g satisfies the contraction condition (2.2.3). 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*

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

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

Proof. 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 (2.2.2) 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 (2.2.3), we have $\mathrm{E}\big|X_t - g_1(0, Y_{t-1})\big| = \mathrm{E}\big|g_1(X_{t-1}, Y_{t-1}) - g_1(0, Y_{t-1})\big| \leq a\mathrm{E}X_{t-1}$. By induction, it follows that for any $k \geq 1$,

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

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 (2), denote $\mathcal{F}_{k,l}^Y = \sigma\{Y_k, \ldots, Y_l\}$ 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

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

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 $n \geq 1$, the coefficients of absolute regularity become

$$\begin{aligned} \beta(n) &= \mathrm{E}\Big\{ \sup_{A \in \mathcal{F}_{n,\infty}^Y} \big|P(A|\sigma\{X_1, Y_0, Y_{-1}, \ldots\}) - P(A)\big|\Big\} \\ &= \mathrm{E}\Big\{ \sup_{A \in \mathcal{F}_{n,\infty}^Y} \big|P(A|X_1) - P(A)\big|\Big\}. \end{aligned} \tag{2.2.7}$$

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

$$\beta(n) = \mathrm{E}\Big\{ \sup_{A \in \mathcal{B}^\infty} \big|P\big((Y_n, Y_{n+1}, \ldots) \in A|X_1\big) - P\big((Y_n, Y_{n+1}, \ldots) \in A\big)\big|\Big\}. \tag{2.2.8}$$

We will provide an upper bound for (2.2.8) 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 $\{U_k, k \in \mathbb{Z}\}$ as be an iid sequence of uniform $(0,1)$ random variables, and construct the chains as follows:

$$X_n' = g\big(X_{n-1}', F^{-1}_{X_{n-1}'}(U_{n-1})\big), \quad Y_n' = F^{-1}_{X_n'}(U_n),$$
$$X_n'' = g\big(X_{n-1}'', F^{-1}_{X_{n-1}''}(U_{n-1})\big), \quad Y_n'' = F^{-1}_{X_n''}(U_n).$$

Since X_1' and X_1'' are independent, so 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

$$\begin{aligned} & \big|P\big((Y_n, Y_{n+1}, \ldots) \in A | X_1 = x\big) - P\big((Y_n, Y_{n+1}, \ldots) \in A\big)\big| \\ = \; & \big|P\big((Y_n', Y_{n+1}', \ldots) \in A | X_1' = x\big) - P\big((Y_n'', Y_{n+1}'', \ldots) \in A | X_1' = x\big)\big| \\ \leq \; & P\big((Y_n', Y_{n+1}', \ldots) \neq (Y_n'', Y_{n+1}'', \ldots) | X_1' = x\big). \end{aligned} \tag{2.2.9}$$

Therefore the coefficients of absolute regularity are bounded by

$$\beta(n) \leq P\big((Y_n', Y_{n+1}', \ldots) \neq (Y_n'', Y_{n+1}'', \ldots)\big) \leq \sum_{k=0}^{\infty} P(Y_{n+k}' \neq Y_{n+k}''). \tag{2.2.10}$$

Observe that the construction of the two chains agrees with that of geometric moment contracting condition in Definition 6.3.1, so it follows from Proposition 2.2.3 that $\mathrm{E}|X_n' - X_n''| \leq (a+b)^n$ for all n. Then

$$\begin{aligned} P(Y_n' \neq Y_n'') &= \mathrm{E}\{P(Y_n' \neq Y_n'' | X_n, X_n'')\} = \mathrm{E}\{P(|Y_n' - Y_n''| \geq 1 | X_n, X_n'')\} \\ &\leq \mathrm{E}\{E|Y_n' - Y_n''|\big|X_n', X_n'')\} = \mathrm{E}|X_n' - X_n''| \leq (a+b)^n. \end{aligned}$$

Hence according to (2.2.10), 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 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 $\{X_t\}$ is also ergodic. □

When Y_t has a continuous distribution, geometric ergodicity of $\{X_t\}$ can be established under stronger conditions on g. The proof of the result relies on Theorem 6.2.3 since $\{X_t\}$ is ϕ-irreducible due to the continuity of the distribution in this situation.

Proposition 2.2.5. *Assume model (2.2.2) where the support of Y_t is $[0, \infty)$, and that the function g satisfies the contraction condition (2.2.3). Moreover if g is increasing and continuous in (x, y), then*

1. *There exists $g_\infty : [0, \infty)^\infty \to [0, \infty)$ such that $X_t = g_\infty(Y_{t-1}, Y_{t-2}, \ldots)$ almost surely.*

2. *The Markov chain $\{X_t, t \geq 1\}$ is geometrically ergodic provided that $a + b < 1$, and hence $\{(X_t, Y_t)\}$ is stationary and ergodic.*

Proof. (1) follows from the same argument as in the proof of Proposition 2.2.4. As for (2), for any fixed $\epsilon > 0$, define ϕ as Lebesgue measure on $[x^*, \infty)$, where $x^* = (g(0,0) + b\epsilon)/(1-a)$, and let A be a set with $\phi(A) > 0$. To prove the $\phi-$irreducible, we need to show that for any $x_1 \in S$, there exists $n \geq 1$, such that $P^n(x_1, A) > 0$. If $x_1 < x^*$, then $g(x_1, \epsilon) < g(0,0) + ax_1 + b\epsilon \leq x^*$, which implies that $\phi\big(A \cap [g(x_1, \epsilon), \infty)\big) > 0$. Because of the assumptions on the function g, and the fact that the distribution of Y_1 given $X_1 = x_1$ has positive probability everywhere, so $P(x_1, A) > 0$. On the other hand, if $x_1 \geq x^*$, it is easy to see that $g(x_1, \epsilon/2) \leq g(x_1, \epsilon) \leq x_1$. If $g(x_1, \epsilon/2) < x^*$, then by the same argument above, we have $P(x_1, A) > 0$. However, if $g(x_1, \epsilon/2) \geq x^*$, then $ag(x_1, \epsilon/2) + b\epsilon \leq g(x_1, \epsilon/2) - g(0,0) \leq ax_1 + b\epsilon/2$. Hence we have $x^* \leq g(x_1, \epsilon/2) \leq x_1 - (b\epsilon)/(2a)$. By induction, there exists $n \geq 1$ such that

$g(x_n, \epsilon/2) \leq x_1 - n(b\epsilon)/(2a) < x^*$, where $x_t = g(x_{t-1}, \epsilon/2)$ for $t = 1, \ldots, n$. Since $\epsilon > 0$, and the function g is increasing in both coordinates, so $P^{n+1}(x_1, A) > 0$. Hence $\{X_t, t \geq 1\}$ is $\phi-$irreducible.

We now show that $\{X_t, t \geq 1\}$ is aperiodic, i.e., a $\phi-$irreducible Markov chain is said to be aperiodic if there exists a small set A with $\phi(A) > 0$ such that for any $x \in A$, $P(x, A) > 0$ and $P^2(x, A) > 0$. Note that in the setting of the proposition, any compact set is a small set. So we take $A = [x^*, K]$ for some positive K large enough. For any $x_1 \in A$, from the proof of $\phi-$irreducibility, it is easy to see that $P(x_1, A) > 0$. Similarly we have $P^2(x, A) = P(X_2 \in A | X_0 = x) \geq P(X_2 \in A | X_1 \in A) P(X_1 \in A | X_0 = x) > 0$.

To check the drift condition, let $V(x) = 1 + x$. There exists $\delta > 0$, such that $a + b < 1 - \delta$. For $x \geq (g(0,0) + \delta)/(1 - a - b - \delta)$, we have

$$\begin{aligned} \mathrm{E}\{V(X_1) | X_0 = x\} &= \mathrm{E}(1 + X_1 | X_0 = x) = 1 + \mathrm{E}\{g(x, Y_0) | X_0 = x\} \\ &\leq 1 + g(0,0) + (a+b)x \leq (1-\delta)(1+x) = (1-\delta)V(x). \end{aligned}$$

Hence the drift condition holds by taking the small set $A = [x_0^*, \{g(0,0) + \delta\}/(1 - a - b - \delta)]$, which establishes the geometric ergodicity of $\{X_t\}$. It is well known that a geometrically ergodic Markov chain starting from its stationary distribution is strongly mixing with geometrically decaying rate (see Remark 6.2.1), hence is an ergodic stationary time series. Denote $\{\xi_t, t \geq 1\}$ as a sequence of iid uniform $(0, 1)$ random variables, then it follows from $Y_t = F_{X_t}^{-1}(\xi_t)$ that $\{Y_t, t \geq 1\}$ is stationary and ergodic. □

2.3 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 $\theta_0 = (\theta_1^0, \ldots, \theta_d^0)^T$. Then the likelihood function of model (2.2.2) conditioned on η_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 θ, is given by

$$l(\theta) = \sum_{t=1}^{n} l_t(\theta) = \sum_{t=1}^{n} \{\eta_t(\theta)Y_t - A(\eta_t(\theta))\}, \tag{2.3.1}$$

with score function

$$S_n(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n} \{Y_t - B(\eta_t(\theta))\}\frac{\partial \eta_t(\theta)}{\partial \theta}. \tag{2.3.2}$$

The maximum likelihood estimator $\hat{\theta}_n$ is a solution to the equation $S_n(\theta) = 0$. Let P_{θ_0} be the probability measure under the true parameter θ_0 and unless otherwise indicated, E[$\cdot$] is taken under θ_0. Recall that $X_t = g_\infty^\theta(Y_{t-1}, Y_{t-2}, \ldots)$ according to part (1) of Propositions 2.2.4 and 2.2.5. We will 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 according to the lemma below, which is adapted from Lemma 3.11 in Pfanzagl (1969).

Lemma 2.3.1. *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:*

1. *$\theta \mapsto f_\theta(\mathbf{x})$ is upper-semicontinuous for all $\mathbf{x} \in \mathbb{R}^\infty$.*
2. *$\sup_{\theta\in C} f_\theta(\mathbf{x})$ is Borel measurable for any compact set $C \subset \Theta$.*
3. *$E\{\sup_{\theta\in\Theta} f_\theta(X)\} < \infty$ 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*

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

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 2.3.1, we can derive the strong consistency of the estimates.

Theorem 2.3.1. *Assume model (2.2.2) with the function g satisfying the contraction condition (2.2.3), and that assumptions (A0)-(A5) hold. Then the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent, that is,*

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

Proof. We first show the identifiability and then establish the consistency result using Lemma 2.3.1. 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 $\eta_t(\theta) \geq B^{-1}(x^*_\theta)$. So it follows from assumptions (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 $\mathrm{E}l^+_t(\theta) < \infty$. Denote $M_n(\theta) = \sum_{t=1}^n l_t(\theta)/n$, 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) pp. 284 and 495). In order to prove the identifiability, we need to show that θ_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

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

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. 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

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

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

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

The function f_θ in Lemma 2.3.1 can be defined as

$$f_\theta(\mathbf{y})=y_1B^{-1}(g^\theta_\infty(y_0,y_{-1},\ldots))-A(B^{-1}(g^\theta_\infty(y_0,y_{-1},\ldots))),$$

where $\mathbf{y} = (y_1, y_0, y_{-1}, \ldots)$. Hence it follows from assumption (A2) and Lemma 2.3.1 that $M(\theta)$ is upper-semicontinuous and for any compact subset $K \subset \Theta$, we have $\limsup_{n\to\infty} \sup_{\theta\in K} M_n(\theta) \leq \sup_{\theta\in K} M(\theta)$. Take $\mathcal{U}_0$ as a local base of θ_0 and let $U \in \mathcal{U}_0$ be a neighborhood of θ_0, then Lemma 2.3.1 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

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta\setminus U} M_n(\theta) \leq \sup_{\theta\in\Theta\setminus U} M(\theta) < M(\theta_0), \quad P_{\theta_0}\text{-a.s.} \tag{2.3.3}$$

Notice that for any $\tilde{\theta} \notin U$, $M_n(\tilde{\theta}) \leq \sup_{\theta\in\Theta\setminus U} M_n(\theta)$. Let $\omega \in \Omega$ such that (2.3.3) 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

$$\begin{aligned} \liminf_{n\to\infty} M_n(\hat{\theta}_n) &\leq \liminf_{n\to\infty, n\in\widetilde{\mathbb{N}}} M_n(\hat{\theta}_n) \leq \limsup_{n\to\infty, n\in\widetilde{\mathbb{N}}} M_n(\hat{\theta}_n) \\ &\leq \limsup_{n\to\infty, n\in\widetilde{\mathbb{N}}} \sup_{\theta\notin U} M_n(\theta) \leq \limsup_{n\to\infty} \sup_{\theta\notin U} M_n(\theta). \end{aligned} \tag{2.3.4}$$

However, according to (2.3.3), we have

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

which contradicts (2.3.4). 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. □

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, η_t and $\dot{\eta}_t$ are both evaluated at θ_0, i.e., $\eta_t = \eta_t(\theta_0)$ and $\dot{\eta}_t = (\partial\eta_t/\partial\theta)|_{\theta=\theta_0}$.

Theorem 2.3.2. *Assume model (2.2.2) with the function g satisfying the contraction condition (2.2.3), and that assumptions (A0)-(A7) hold. Then the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically normal, i.e.,*

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

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

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

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

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

$$\begin{aligned}
R_n^\dagger(u) &= l^\dagger(\theta_0) - l^\dagger(\theta_0 + un^{-1/2}) \\
&= \sum_{t=1}^{n} Y_t\eta_t - \sum_{t=1}^{n} A(\eta_t) - \sum_{t=1}^{n}(\eta_t + u^T n^{-1/2}\dot{\eta}_t)Y_t + \sum_{t=1}^{n} A(\eta_t + u^T n^{-1/2}\dot{\eta}_t) \\
&= -u^T n^{-1/2} \sum_{t=1}^{n} Y_t\dot{\eta}_t + \sum_{t=1}^{n}\{A(\eta_t + u^T n^{-1/2}\dot{\eta}_t) - A(\eta_t)\} \\
&= -u^T n^{-1/2} \sum_{t=1}^{n}\{Y_t - B(\eta_t)\}\dot{\eta}_t \\
&\quad + \sum_{t=1}^{n}\{A(\eta_t + u^T n^{-1/2}\dot{\eta}_t) - A(\eta_t) - u^T n^{-1/2} B(\eta_t)\dot{\eta}_t\}. \qquad (2.3.5)
\end{aligned}$$

Let $s_t = n^{-1/2}\{Y_t - B(\eta_t)\}\dot{\eta}_t$, 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. Note that

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

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

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

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 (2.3.5) 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 $R_n^{\dagger}(u) \xrightarrow{\mathcal{L}} -u^T V + \frac{1}{2} u^T \Omega u$, where $V \sim N(0, \Omega)$. It then follows that $\mathrm{argmin}_u R_n^{\dagger}(u) \xrightarrow{\mathcal{L}} \mathrm{argmin}_u \{-u^T V + \frac{1}{2} u^T \Omega u\} = \Omega^{-1} V \sim N(0, \Omega^{-1})$.

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

the difference becomes

$$\begin{aligned} R_n^{\dagger}(u) - R_n(u) &= \sum_{t=1}^{n}\{Y_t - B(\eta_t)\}\{\eta_t(\theta) - \eta_t - u^T n^{-1/2}\dot{\eta}_t\} \\ &\quad - \sum_{t=1}^{n}[A(\eta_t(\theta)) - A(\eta_t + u^T n^{-1/2}\dot{\eta}_t) \\ &\quad - B(\eta_t)\{\eta_t(\theta) - \eta_t - u^T n^{-1/2}\dot{\eta}_t\}]. \end{aligned} \tag{2.3.6}$$

By Taylor expansion, the first term in (2.3.6) 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

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

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, so the first term in (2.3.6) converges to 0 uniformly on $[-K, K]$ for any $K > 0$. We now apply Taylor expansion to each component in the second term of (2.3.6),

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

where $0 \le c \le u^T n^{-1/2}\dot{\eta}_t$, θ_1^* and θ_2^* both lie between θ_0 and θ. Therefore the second

term in (2.3.6) becomes

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

which converges to 0 on a compact set of u under smoothness assumptions. So (2.3.6) converges to 0 as $n \to \infty$, which implies that $\text{argmin}_u R_n(u)$ and $\text{argmin}_u R_n^\dagger(u)$ have the same asymptotic distribution, i.e.,

$$\text{argmin}_u R_n(u) \xrightarrow{\mathcal{L}} \Omega^{-1}V \sim N(0, \Omega^{-1}).$$

Note that $\text{argmin}_u R_n(u) = \text{argmax}_u\ l(\theta_0 + un^{-1/2}) = \sqrt{n}(\hat{\theta}_n - \theta_0)$, where $\hat{\theta}_n$ is the conditional maximum likelihood estimator. Hence

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

□

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

2.4 Examples

2.4.1 Linear Dynamic Models

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

$$Y_t|\mathcal{F}_{t-1} \sim p(y|\eta_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1}, \tag{2.4.1}$$

where $X_t = B(\eta_t) = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$, and $\delta > 0, \alpha, \beta \geq 0$ are parameters. Observe that model (2.4.1) is a special case of model (2.2.2) by defining the function g_θ as

$$g_\theta(x, y) = \delta + \alpha x + \beta y, \tag{2.4.2}$$

with $\theta = (\delta, \alpha, \beta)^T$ and the contraction condition (2.2.3) corresponds to $\alpha + \beta < 1$. Note that by recursion we have, for all t,

$$X_t(\theta) = \delta/(1-\alpha) + \beta \sum_{k=0}^{\infty} \alpha^k Y_{t-1-k}. \tag{2.4.3}$$

It follows that $X_t(\theta) \geq x^* = \delta/(1-\alpha)$ since Y_t only takes non-negative values. A direct application of Propositions 2.2.3, 2.2.4 and 2.2.5 gives the stability properties of model (2.4.1).

Proposition 2.4.1. *Assume model (2.4.1) with $\alpha+\beta < 1$. Then the process $\{X_t, t \geq 1\}$ 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 model (2.4.1) are given by (2.3.1) and (2.3.2) 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 re-

cursively by

$$\frac{\partial \eta_t}{\partial \theta} = \begin{pmatrix} 1 \\ B(\eta_{t-1}) \\ Y_{t-1} \end{pmatrix} / B'(\eta_t) + \alpha \frac{B'(\eta_{t-1})}{B'(\eta_t)} \frac{\partial \eta_{t-1}}{\partial \theta}. \tag{2.4.4}$$

The maximum likelihood estimator $\hat{\theta}_n$ is a solution of the equation $S_n(\theta) = 0$. Furthermore, the Hessian matrix can be found by taking derivatives of the score function, i.e.,

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

where

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

It follows from the representation with the infinite past (2.4.3) that assumptions (A1)-(A3) and (A6) are satisfied. In order to apply Theorem 2.3.2 when investigating the asymptotic behavior of the MLE, we need to impose the following regularity conditions:

(L0) The true parameter vector 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 2.4.1. *Assume model (2.4.1) and that assumptions (L0)-(L2) hold. Then the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent and asymptotically normal, i.e.,*

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

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

Proof. According to Theorems 2.3.1 and 2.3.2, 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_{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 (2.4.3) 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 $Y_{t-1} \in \text{span}\{Y_{t-2}, Y_{t-3}, \ldots\}$ which contradicts the fact that $\text{Var}(Y_{t-1} |\mathcal{F}_{t-2}) > 0$. So 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. □

Remark 2.4.1. Under the contraction condition $\alpha + \beta < 1$, $\{Y_t\}$ can be represented as a causal ARMA(1,1) process. To see this, denote $d_t = Y_t - X_t$, then it follows from $\text{E}(d_t|\mathcal{F}_{t-1}) = 0$ that $\{d_t, t \in \mathbb{Z}\}$ is a martingale difference sequence. Therefore model (2.4.1) can be written as

$$Y_t - (\alpha + \beta)Y_{t-1} = \delta + d_t - \alpha d_{t-1}. \tag{2.4.5}$$

Denote $\gamma_Y(h)$ as 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).

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.

Remark 2.4.2. A sufficient condition for assumption (L1) is

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

provided that $\delta_U/\epsilon + \sum_{k=1}^{\infty}(1-\epsilon)^k Y_{1-k}$ is in the range of $B(\eta)$. This can be seen by noting that $X_1(\theta) \le \delta_U/\epsilon + \sum_{k=1}^{\infty}(1-\epsilon)^k Y_{1-k}$.

Remark 2.4.3. If $A''(\eta_t) \ge \underline{c}$ for some $\underline{c} > 0$, 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 $\gamma_Y(0) < \infty$.

Proof. 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 $\partial B(\eta_1)/\partial\alpha = \delta/(1-\alpha)^2 + \beta\sum_{k=1}^{\infty} k\alpha^{k-1}Y_{-k}$. Then 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 $\gamma_Y(0) < \infty$. □

Next we consider some specific models belonging to class (2.4.1), most of which are geared towards modeling time series of counts.

Example 2.4.1. As a special case of the linear dynamic model (2.4.1) with $\eta_t = \log\lambda_t$ and $A(\eta_t) = e^{\eta_t}$, the Poisson INGARCH$(1,1)$ model is given by

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{Pois}(\lambda_t), \quad \lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1}, \tag{2.4.6}$$

where $\delta > 0, \alpha, \beta \geq 0$ are parameters. According to Proposition 2.4.1, it is easy to see that if $\alpha + \beta < 1$, then $\{\lambda_t\}$ is geometric moment contracting and has a unique stationary distribution π; moreover if $\lambda_1 \sim \pi$, then $\{(Y_t, \lambda_t), t \geq 1\}$ is an ergodic stationary process. As for inference, the MLE $\hat{\theta}_n$ is strongly consistent and asymptotically normal according to Theorem 2.4.1, i.e., $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega^{-1})$, as $n \to \infty$, where $\Omega = \mathrm{E}\{1/\lambda_t(\partial\lambda_t/\partial\theta)(\partial\lambda_t/\partial\theta)^T\}$. 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 $\mu = \mathrm{E}Y_t = \delta/(1-\alpha-\beta)$ and $C(\theta)$ is a positive constant dependent on θ. Hence by monotone convergence theorem, we have

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

Hence assumption (L1) holds according to Remark 2.4.2. Notice that $B(\eta_t) = \lambda_t \geq \lambda^* := \delta/(1-\alpha)$ for all t, so $A''(\eta_t) = e^{\eta_t}$ is bounded away from 0, so assumption (L2) holds according to Remark 2.4.3.

Example 2.4.2. The negative binomial INGARCH$(1,1)$ model (NB-INGARCH) is defined as

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{NB}(r, p_t), \quad X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1}, \tag{2.4.7}$$

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

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

When $r = 1$, the conditional distribution of Y_t becomes geometric distribution with probability of success p_t, in which case (2.4.7) reduces to a geometric INGARCH model.

By virtue of Proposition 2.4.1, if $\alpha + \beta < 1$, then $\{X_t, t \geq 1\}$ is a geometric moment contracting Markov chain, and has 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 which maximizes the profile likelihood, and thus $\hat{\theta}$ can be otained correspondingly. 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 θ_0 and covariance matrix Ω^{-1}/n, where $\Omega = \mathrm{E}\{r/X_t/(X_t + r)(\partial X_t/\partial\theta)(\partial X_t/\partial\theta)^T\}$. 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 2.4.2. Note that $A''(\eta_t) = re^{\eta_t}/(1-e^{\eta_t})^2$ is increasing, so assumption (L2) holds provided $\gamma_Y(0) < \infty$ according to Remark 2.4.3. Because $\mathrm{Var}(X_1) = \alpha^2\mathrm{Var}(X_0) + \beta^2\mathrm{Var}(Y_0) + 2\alpha\beta\mathrm{Cov}(X_0, Y_0)$, where

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

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

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

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

Example 2.4.3. We define the binomial INGARCH$(1,1)$ model as

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{B}(m, p_t), \quad mp_t = \delta + \alpha mp_{t-1} + \beta Y_{t-1}, \tag{2.4.8}$$

where $\delta > 0, \alpha, \beta \geq 0$ are parameters and $\delta + \alpha m + \beta m \leq m$ since $p_t \in (0,1)$. This implies the contraction condition $\alpha + \beta < 1$. In particular, when $m = 1$, it models time series of binary data, and is called a Bernoulli INGARCH model. If $\delta + \alpha m + \beta m \leq m$, then $\{X_t = mp_t, t \geq 1\}$ is geometric moment contracting and has a unique stationary distribution π; furthermore, $\{(X_t, Y_t), t \geq 1\}$ is ergodic when $X_1 \sim \pi$.

We now consider the inference of the model. Firstly, because of the special constraint $p_t \in (0,1)$, the parameter space becomes

$$\Theta = \{(\delta, \alpha, \beta)^T : 0 < \delta_L \leq \delta \leq \delta_U, \epsilon \leq \alpha + \beta \leq 1 - \epsilon\} \text{ for some } \epsilon > \delta_U/m.$$

Since $Y_t \leq m$, so $X_1(\theta) \leq (\delta + \alpha m)/(1-\alpha)$ and $B^{-1}(X_1(\theta)) \leq \log\{(\delta_U + (1-\epsilon)m)/(\epsilon m - \delta_U)\}$. Hence assumption (L1) holds. Notice that $A''(\eta_t) = mp_t(1-p_t)$ and $p_t \in [\delta_U/m, (\delta + \beta m)/(m(1-\alpha))] \subsetneq [0,1]$, so $A''(\eta_t)$ is bounded away from 0. Similar to the proof in Example 2.4.2, one can show that $\gamma_Y(0) < \infty$ provided that $(\alpha+\beta)^2 + \beta^2/m < 1$. So assuming m is known and $(\alpha+\beta)^2 + \beta^2/m < 1$, the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent and asymptotically normal with mean θ_0 and covariance matrix Ω^{-1}/n, where $\Omega = \mathrm{E}\{m/X_t/(m - X_t)(\partial X_t/\partial\theta)(\partial X_t/\partial\theta)^T\}$.

Example 2.4.4. The gamma INGARCH model, which has a continuous response, is given by

$$Y_t|\mathcal{F}_{t-1} \sim \Gamma(\kappa, s_t), \quad s_t = \delta/\kappa + \alpha s_{t-1} + \beta/\kappa Y_{t-1}, \tag{2.4.9}$$

where κ and s_t are the shape and scale parameters of the gamma distribution respectively and $\delta > 0, \alpha, \beta \geq 0$ are parameters. Here the natural parameter is $\eta_t = -1/s_t$ and the Markov chain $X_t = B(\eta_t) = -\kappa/\eta_t$. If $\alpha + \beta < 1$, then $\{X_t = \kappa s_t, t \geq 1\}$ is geometric moment contracting and has a unique stationary distribution π; furthermore, $\{(Y_t, X_t), t \geq 1\}$ is an ergodic stationary process if $X_1 \sim \pi$.

As for the inference in this model, assume κ is known and $(\alpha+\beta)^2+\beta^2/\kappa<1$. Then the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent and asymptotically normal with mean θ_0 and covariance matrix Ω^{-1}/n where $\Omega=\mathrm{E}\{\kappa/s_t^2(\partial s_t/\partial\theta)(\partial s_t/\partial\theta)^T\}$. To see this, note that $B^{-1}(x)=-\kappa/x<0$ when $x>0$, which verifies assumption (L1) according to Remark 2.4.2. Similar to the proof in Example 2.4.2, one can show that $\gamma_Y(0)=(1/\kappa+1)\gamma_X(0)+\mu^2/\kappa$ and $\gamma_X(0)=(\beta^2\mu^2/\kappa)/\{1-(\alpha+\beta)^2-\beta^2/\kappa\}$. Hence as long as $(\alpha+\beta)^2+\beta^2/\kappa<1$, we have $\gamma_Y(0)<\infty$. Since $A''(\eta_t)=\kappa/\eta_t^2\geq\delta_L^2/\kappa>0$, assumption (L2) holds according to Remark 2.4.3.

2.4.2 Nonlinear Dynamic Models

It is possible to generalize (2.4.1) to nonlinear dynamic models. One approach 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

$$Y_t|\mathcal{F}_{t-1}\sim p(y|\eta_t),\quad X_t=\delta+\alpha X_{t-1}+\beta Y_{t-1}+\sum_{k=1}^{K}\beta_k(Y_{t-1}-\xi_k)^+, \qquad (2.4.10)$$

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 $K=0$, (2.4.10) reduces to the linear model (2.4.1). It is easy to see that model (2.4.10) is a special case of model (2.2.2) by defining $g_\theta(x,y)=\delta+\alpha x+\beta y+\sum_{k=1}^K\beta_k(y-\xi_k)^+$, where $\theta=(\delta,\alpha,\beta,\beta_1,\ldots,\beta_K)^T$. Note that in each of the pieces segmented by the knots, (2.4.10) 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)). According to Propositions 2.2.3, 2.2.4 and 2.2.5, we can establish the stability properties of the model.

Proposition 2.4.2. *Consider model (2.4.10) with parameters satisfying* $\alpha + \beta < 1, \beta + \sum_{k=1}^{s} \beta_k \geq 0$ *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 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 (2.3.1). 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,

$$\begin{aligned} X_t &= \delta/(1-\alpha) + \beta \sum_{i=0}^{\infty} \alpha^i Y_{t-1-i} + \sum_{k=1}^{K} \beta_k \sum_{i=0}^{\infty} \alpha^i (Y_{t-1-i} - \xi_k)^+ \\ &= \delta/(1-\alpha) + \sum_{i=0}^{\infty} \alpha^i \{\beta Y_{t-1-i} + \sum_{k=1}^{K} \beta_k (Y_{t-1-i} - \xi_k)^+\}. \end{aligned} \tag{2.4.11}$$

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 in order to apply Theorem 2.4.1, we only need to impose the following regularity assumptions for the nonlinear model (2.4.10):

(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 2.4.2.

(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 2.4.2 and 2.4.3. The asymptotic properties of the MLE are summarized in the following theorem.

Theorem 2.4.2. *For model (2.4.10), 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.,*

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

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

Proof. According to Theorem 2.3.2, we only need to establish the identifiability of the model. Similar to the proof of Theorem 2.4.1, 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

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

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. □

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

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{Pois}(\lambda_t), \quad \lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1} + \sum_{k=1}^{K} \beta_k (Y_{t-1} - \xi_k)^+. \qquad (2.4.12)$$

It follows that under the conditions of Proposition 2.4.2 and Theorem 2.4.2 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

$$\frac{\partial \lambda_t}{\partial \theta} = \begin{pmatrix} 1 & \lambda_{t-1} & Y_{t-1} & (Y_{t-1} - \xi_1)^+ & \dots & (Y_{t-1} - \xi_K)^+ \end{pmatrix}^T + \alpha \frac{\partial \lambda_{t-1}}{\partial \theta}.$$

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

$$Y_t | \mathcal{F}_{t-1} \sim \text{Pois}(\lambda_t), \quad \lambda_t = (\alpha_0 + \alpha_1 \exp\{-\gamma \lambda_{t-1}^2\})\lambda_{t-1} + \beta Y_{t-1}, \tag{2.4.13}$$

where $\alpha_0, \alpha_1, \beta, \gamma > 0$ are parameters. We point out that if $\alpha_0 + \alpha_1 + \beta < 1$, then model (2.4.13) belongs to the class of models (2.2.2) and hence enjoys the stability properties stated in Propositions 2.2.3 and 2.2.4. As for the inference of the model, we refer readers to Fokianos *et al.* (2009) for details.

2.5 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 (2.3.1) 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 (2.4.13) will be applied to two real datasets, and tools for checking goodness of fit will be considered.

2.5.1 Simulation for the Nonlinear Model

As specified in (2.4.12), a 1-knot nonlinear dynamic model is simulated according to

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

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 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 (2.4.10) with only one knot at 5. The means and standard errors of the estimates from all 1000 runs are summarized in Table 2.1 and the histograms of the estimates are depicted in Figure 2.1. The performance of these estimates is reasonably good and consistent with the theory described in Theorem 2.4.2. 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 2.2. Here the locations of the knots are determined by sample quantiles. Not surprisingly, the performance of the maximum likelihood estimates of β and β_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 2.3. 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

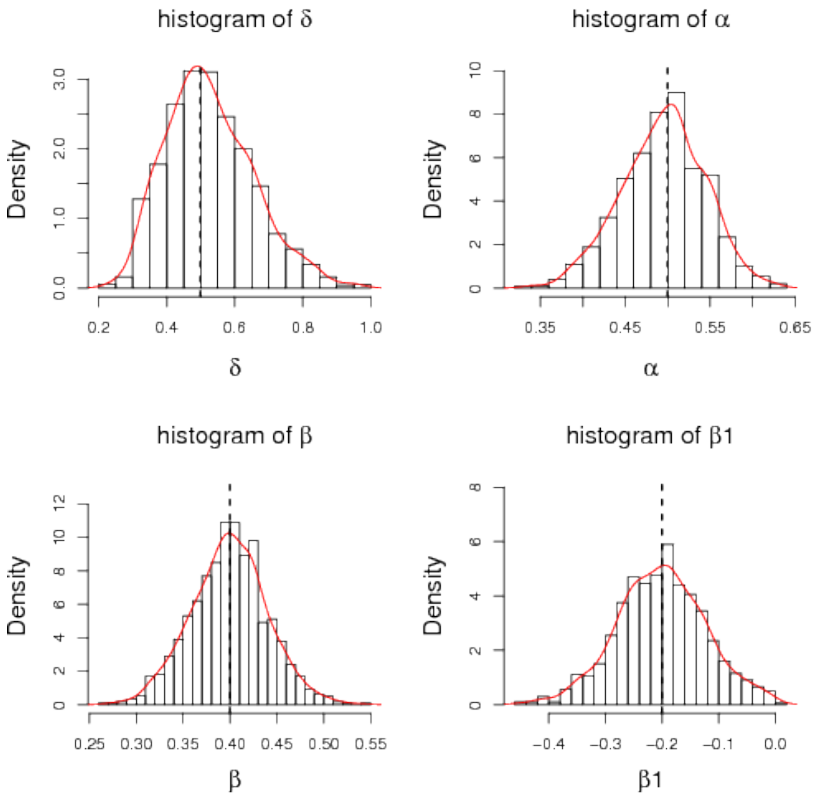

Figure 2.1: Histograms of the 1-knot model with sample size 1000 assuming the knot is known. The overlaying curves are the density estimates and the dashed vertical lines represent the true values of the parameters.

Table 2.1: 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	1000
s.e.	(0.0041)	(0.0016)	(0.0013)	(0.0025)	

Table 2.2: 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	1000
s.e.	(0.0042)	(0.0016)	(0.0015)	(0.0023)	

cases when $n = 1000$. In light of the idea of interpolating the nonlinear dynamic of λ_t by a piecewise linear function, we plot in Figure 2.2 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 2.3: 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%	1000
BIC	59.4%	38.4%	2.0%	0.2%	0	

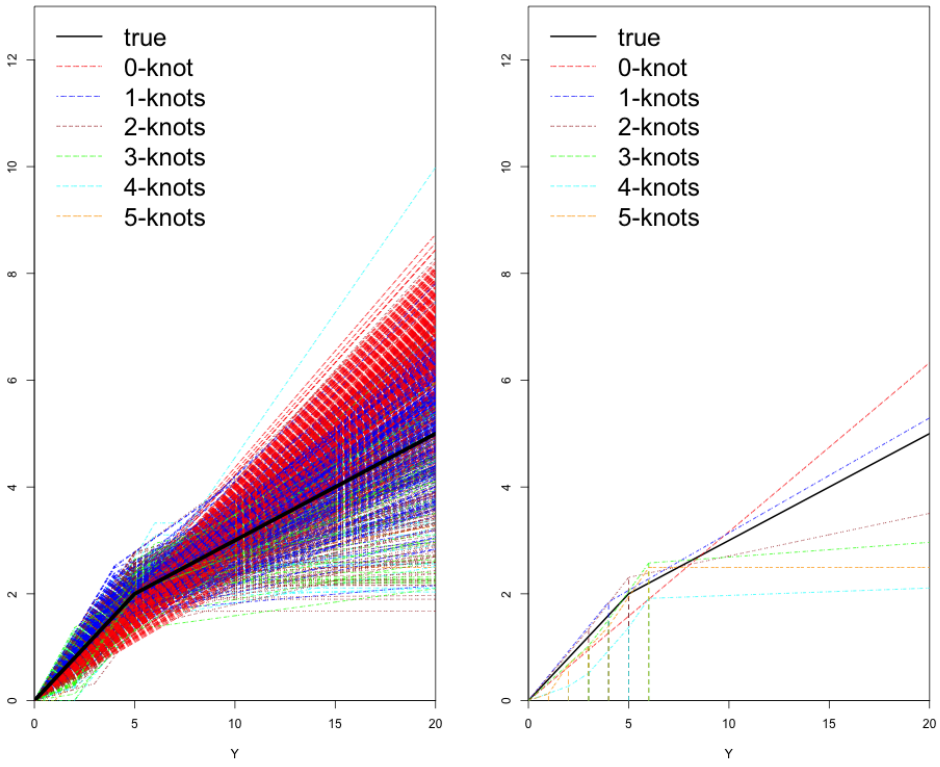

Figure 2.2: Left: the black curve is the true function $0.4y - 0.2(y-5)^+$, and the other curves are the piecewise linear functions fitted in each simulation where the number of knots K is selected via AIC; Right: for each value of K, we plot the fitted curve from one specific run that chooses the particular number of knots.

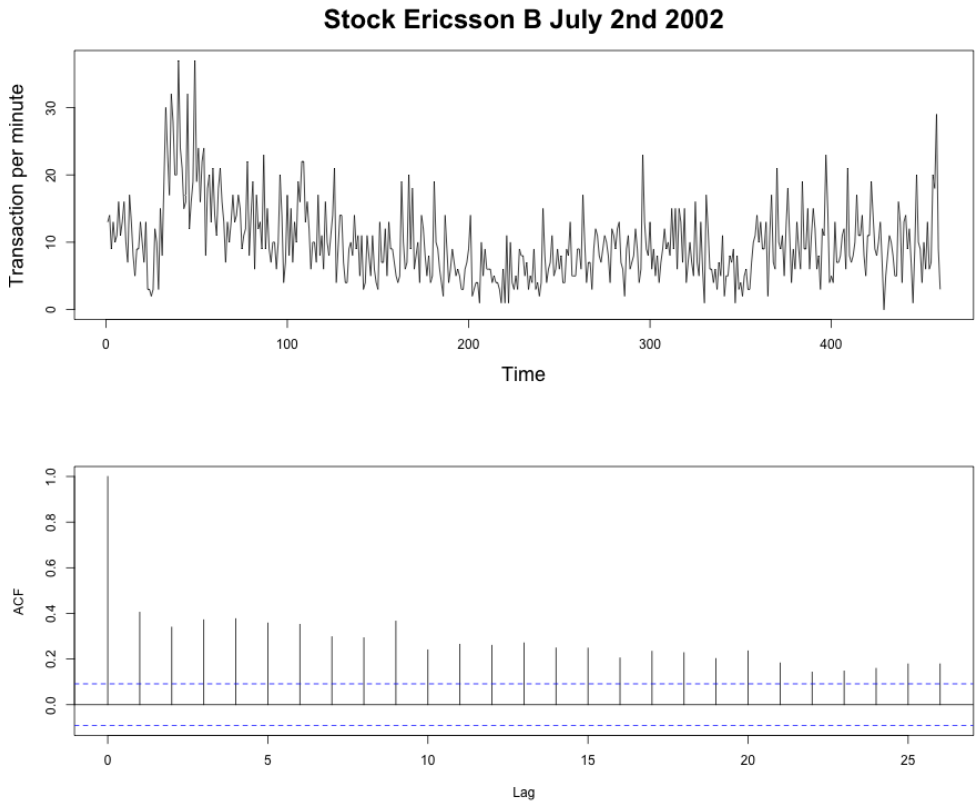

Figure 2.3: Top: Number of transactions per minute of the stock Ericsson B during July 2nd 2002; Bottom: ACF of the data.

2.5.2 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 2.3 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

Table 2.4: 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

is given by

$$\begin{aligned} \hat{\lambda}_t &= 0.2912 + 0.8312\hat{\lambda}_{t-1} + 0.1395Y_{t-1}, \\ &\quad (0.1000)\ (0.0242) \qquad (0.0188) \end{aligned}$$

and the fitted NB-INGARCH model is

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

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 2.3.2.

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 2.4. The fitted 1-knot Poisson model, which has the smallest AIC, is given by

$$\begin{aligned} \hat{\lambda}_t &= 0.5837 + 0.8319\hat{\lambda}_{t-1} + 0.0906Y_{t-1} + 0.0722(Y_{t-1}-9)^+. \\ &\quad (0.1884)\ (0.0241) \qquad (0.0295) \qquad (0.0373) \end{aligned}$$

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

$$\begin{aligned} \hat{X}_t &= 0.4931 + 0.8444\hat{X}_{t-1} + 0.0903Y_{t-1} + 0.0603(Y_{t-1}-9)^+. \\ &\quad (0.2559)\ (0.0350) \qquad (0.0412) \qquad (0.0546) \end{aligned}$$

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 (2.4.13) is also applied to this dataset by Fokianos *et al.* (2009) and is given by

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

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. We first consider the standardized Pearson residuals $e_t = (Y_t - \text{E}(Y_t|\mathcal{F}_{t-1}))/\sqrt{\text{Var}(Y_t|\mathcal{F}_{t-1})}$ which can be obtained by replacing the population quantities by their estimated

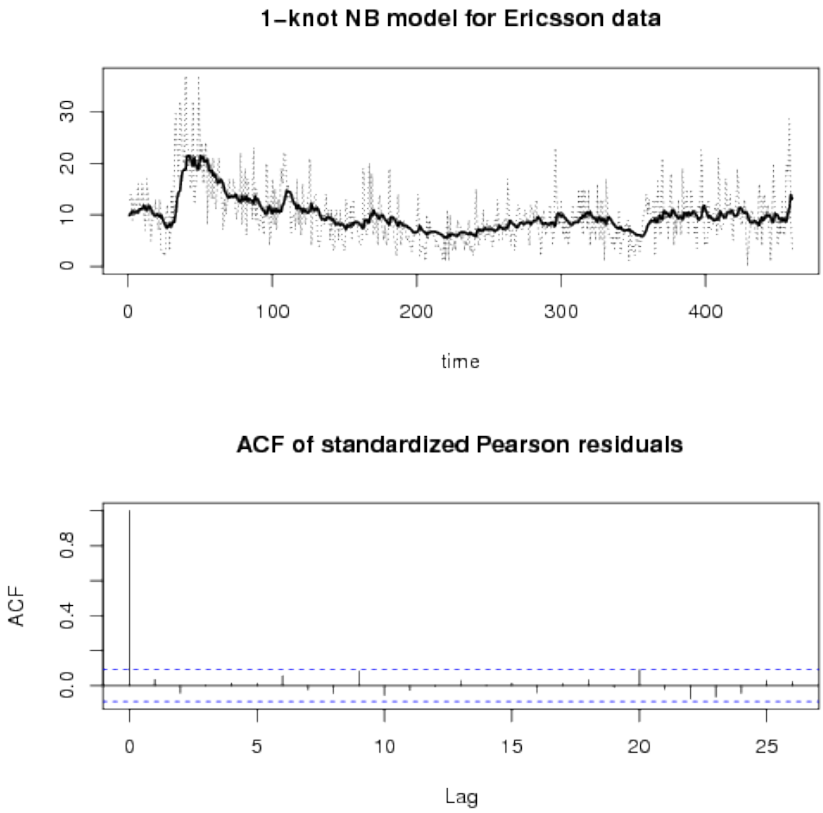

Figure 2.4: Top: Dotted curve represents the number of transactions of Ericsson stock, and the overlaying curve is the fitted conditional mean process by 1-knot NB-based model; Bottom: ACF of the standardized Pearson residuals.

counterparts. If the model is correctly specified, then the residuals $\{\hat{e}_t\}$ 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 2.4 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

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

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 $\tilde{u}_t$ 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 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 2.5, and the p-values are reported in Table 2.5. 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 $(n-1)^{-1}\sum_{t=2}^{n} s(F_t(Y_t))$ where $F_t(\cdot)$ is the CDF of the prediction distribution and $s(\cdot)$ denotes some scoring rule. The first scoring rule we consider is the logarithm

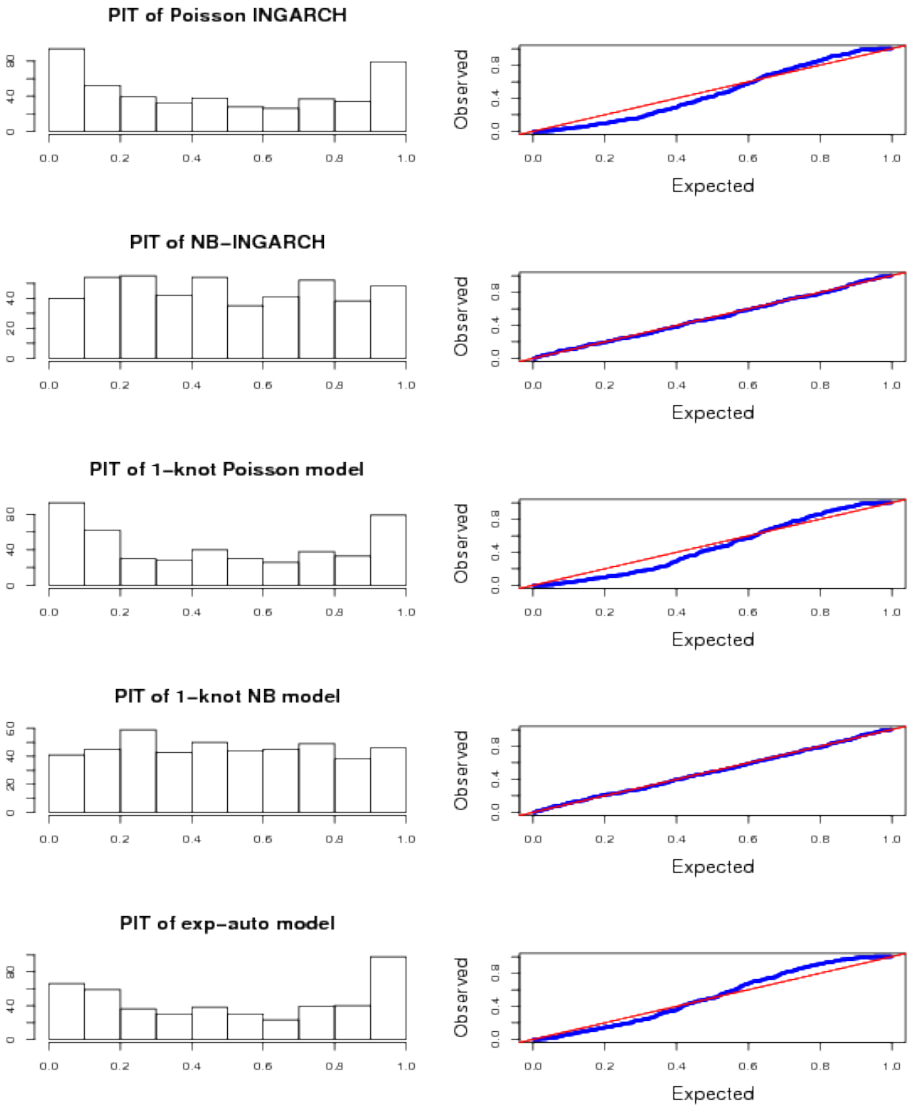

Figure 2.5: Left: histograms of randomized PIT's for all of the models fitted to the Ericsson stock data; Right: QQ-plots of $\tilde{u}_t$ against standard uniform distribution for the corresponding models, where the straight line is the 45° line with zero intercept.

Table 2.5: 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

score (LS), which is closely related to the classical Shannon entropy and is defined as

$$s(F_t(Y_t)) := -\log p_t(Y_t), \tag{2.5.1}$$

where $p_t(Y_t)$ is the probability mass function of the predictive distribution at the observed count at time t. The quadratic score (QS) involves an augmentation of the information collected in the logarithmic score by a summary measure from all probability ordinates, denoted by $\|p_t\|^2 = \sum_{j=0}^{\infty} p_t(j)^2$, and is given by

$$s(F_t(Y_t)) := -2p_t(Y_t) + \|p_t\|^2. \tag{2.5.2}$$

The last score we consider here is the ranked probability score (RPS), which is defined as

$$s(F_t(Y_t)) := \sum_{j=0}^{\infty} \{F_t(j) - \mathbf{1}_{[Y_t \le j]}\}^2. \tag{2.5.3}$$

For details and properties of these scores, readers can refer to Czado *et al.* (2009) and Jung and Tremayne (2011). Table 2.5 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 geometrically distributed with probability of success $p = 0.1$ (Chang (2010)). Figure 2.6 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 2.6 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 (2.4.7) 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

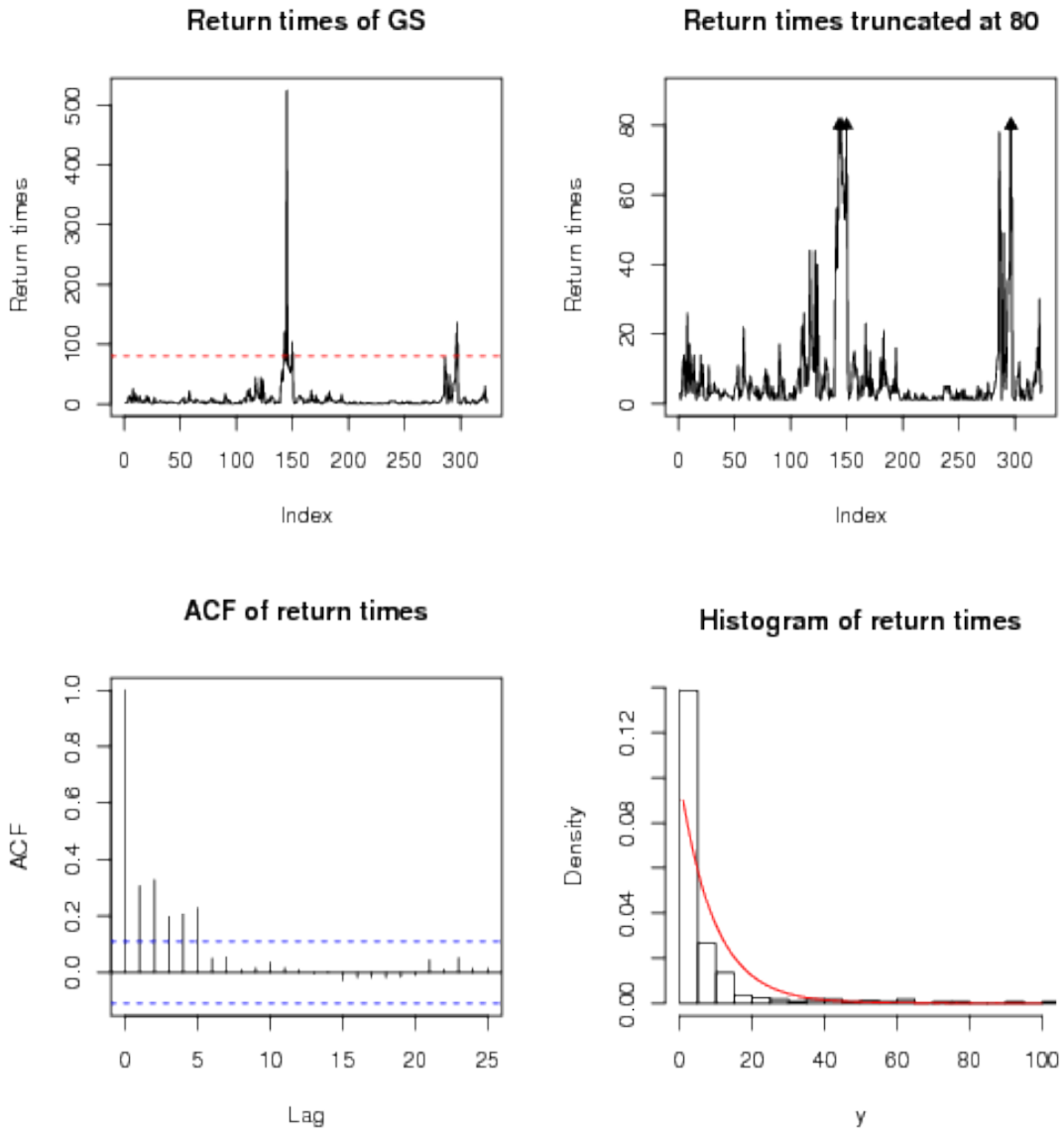

Figure 2.6: Top left: Return times of GS stock, the dashed horizontal line locates at 80; Top right: Return times truncated at 80 in order to ameliorate the visual effect of the five large observations that are represented by solid triangles; Bottom left: ACF of the return times; Bottom right: Histogram of the return times, where the curve overlaid is the density function of a geometric distribution with $p = 0.1$.

Table 2.6: 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**

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 $X_t = (1 - p_t)/p_t$. 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 2.4.2 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 2.7, and the prediction scores and the p-values of the PIT test are summarized in Table 2.6. Two Poisson-based models are also included for comparison, and as expected, they do not perform as well as the geometric-based models.

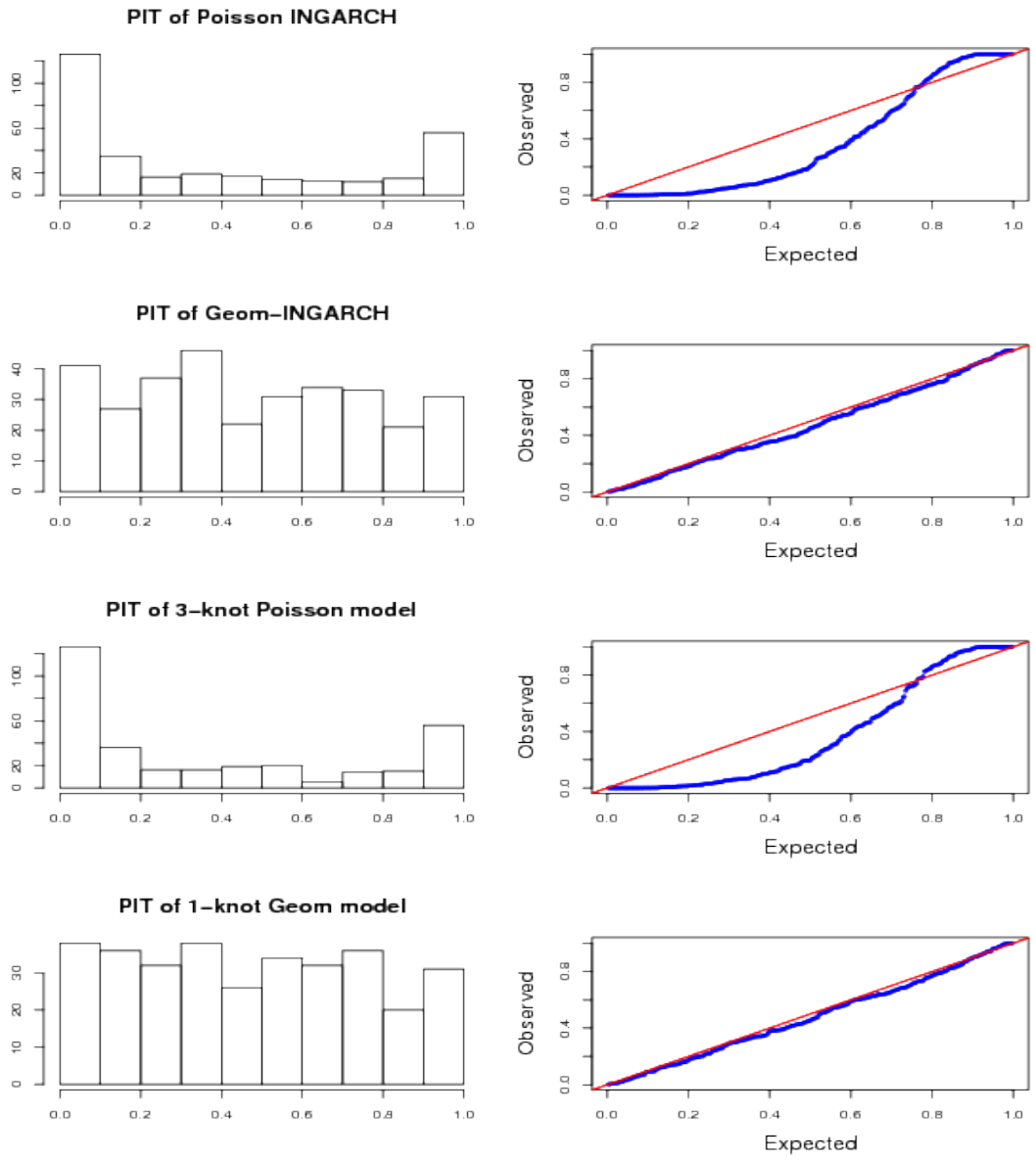

Figure 2.7: Left: histograms of randomized PIT's for the models fitted to GS return times; Right: QQ-plots of $\tilde{u}_t$ against standard uniform distribution for the corresponding models, where the straight line is the 45° line with zero intercept.

Chapter 3

Extensions of INGARCH Models

3.1 Introduction

This chapter aims at extending the broad class of models introduced in the last chapter in several directions. The first extension considers a more generalized INGARCH structure that allows for a threshold in the dynamics of the intensity process. As demonstrated in Fokianos *et al.* (2009) and Davis and Liu (2012), the Poisson INGARCH is incapable of modeling negative serial dependence in the observations. To this end, a self-excited threshold Poisson autoregression, also called a self-excited threshold INGARCH (SETINGARCH), is introduced by adapting the idea of a self-excited threshold ARMA process (Tong (1990)). As a result, this introduces a more general and flexible autoregression modeling framework. It is shown that under some constraints on the parameter space, the count process is stationary and admits a strong law of large numbers. In particular, even when one of the regimes is explosive, the resulting process still has a unique stationary distribution. The estimation procedure and the relevant asymptomatic theory are studied. Two simulation studies are

presented and a real data application is considered.

Another extension is concerned with incorporating explanatory covariates into an INGARCH model. It is of practical and primary interest to investigate regression effects of covariates on a time series of counts. However, it is difficult to do so while maintaining interpretability (see for example Jung and Tremayne (2011)). This in turn complicates the stability properties of the model, which are required for establishing asymptotically correct inference procedure. One approach to tackle this problem, which may have limited applicability in practice, is to consider the covariates as a realization of a special type of a stochastic process. Specifically, it modifies the conditional distribution of the observations and the dynamics of the conditional mean process simultaneously. It is shown that under some conditions on the covariates, the process in the model is strictly stationary and geometrically ergodic. Three real data applications are considered and comparisons between models with and without covariates are provided.

Finally, the class of models proposed in the last chapter is generalized to higher orders, that is, the conditional mean X_t is a function of its own lagged values and the previous observations at lags of any order. By utilizing the IRF approach, stationarity and ergodicity are established for the count process under a contracting constraint on the parameters. The real data application of this model is postponed till the next chapter, where the comparison between a model with lags of higher orders and its bivariate counterpart will be drawn.

The organization of this chapter is as follows: Section 3.2 introduces the self-excited threshold INGARCH model. The relevant stability results, likelihood inference and the corresponding asymptotic theory can be found in Wang *et al.* (2012). A real data application to the frequencies of occurrences of gold particles is given as well. The model with covariates is proposed in Section 3.3, and conditions under

which the process is stationary and geometrically ergodic is provided. Three data applications are considered, including the number of road crashes near the Schiphol airport in the Netherlands, the number of incidences of polio in the US (polio data) and the number of asthma presentations in an Australian hospital (asthma data). In Section 3.4, a model of arbitrary orders is formulated and the corresponding stability theory is proved.

3.2 Self-Excited Threshold Poisson Autoregression

Despite many of the advantages that an INGARCH model enjoys, it is incapable of modeling negative serial dependence in the observations. This can be seen through the fact that $\{Y_t\}$ can be represented as an ARMA$(1,1)$ process with a sequence of martingale differences as innovations and with a positive autoregressive coefficient (see e.g., Davis and Liu (2012)). To this end, this section proposes a self-excited threshold Poisson autoregression process, also known as a self-excited threshold INGARCH (SETINGARCH), which allows for a more general modeling framework for the intensity process and includes the possibility of negative serial dependence in the data. The model assumes a two-regime structure of the conditional mean process $\{\lambda_t\}$ according to the magnitude of the lagged observations. Such an extension to a model with threshold has its own merits, on account of the successful modeling strategy of a self-excited threshold ARMA process introduced by Tong (1990). The stability properties of the model is derived by drawing upon classical Markov chain theory including e-chain (see Definition 6.2.10 and Theorem 6.2.2) and Lyapunov's method (see e.g., Duflo (1997)). An estimation procedure, which gives the maximum likelihood estimates of the parameters, is proposed, and the relevant asymptotic behavior of the parameter estimates is established. A simulation study with two different sets

of parameters is implemented, and the model is applied to a real data set, which consists of frequencies of occurrences of gold particles.

3.2.1 Model Formulation and Stability Theory

For illustration purposes, only a first order self-excited threshold autoregression model is investigated in this chapter. However, the generalization to a higher order model with multiple thresholds is also possible using similarly stylized arguments. Let $Y_1, Y_2, \ldots$ be observations from a model that is defined recursively in the following fashion,

$$Y_t|\mathcal{F}_{t-1} \sim \text{Pois}(\lambda_t),$$

and the conditional intensity process λ_t follows

$$\lambda_t = \begin{cases} d_1 + a_1\lambda_{t-1} + b_1 Y_{t-1}, & \text{if } \ Y_{t-1} \leq r, \\ d_2 + a_2\lambda_{t-1} + b_2 Y_{t-1}, & \text{if } \ Y_{t-1} > r, \end{cases} \tag{3.2.1}$$

where $\mathcal{F}_t = \sigma\{Y_s, s \leq t\}$, $d_1 > 0$ and $d_2, a_1, a_2, b_1, b_2 \geq 0$ are parameters, and $r \in \mathbb{N} = \{0, 1, \ldots\}$ is the *threshold parameter*.

Let $\theta^{(i)} = (d_i, a_i, b_i), i = 1, 2$ be the regime-specific parameter vector, where $\theta^{(1)} \neq \theta^{(2)}$. The dynamics of the process are governed by a two-regime scheme. In the following context, Y_t is said to lie in the *lower regime* if $Y_t \leq r$, denoted by $Y_t \in R_1$, where $R_1 = \{0, 1, \ldots, r\}$; otherwise, Y_t is said to be in the *upper regime*, denoted by $Y_t \in R_2$, where $R_2 = \mathbb{N}/R_1$. The primary focus of this section is on the conditional mean process $\{\lambda_t\}$, which can be easily seen as a time-homogeneous Markov chain, while the count process $\{Y_t\}$ is not a Markov chain itself.

As was noted by Fokianos *et al.* (2009), it is sometimes easier to work with an underlying Poisson process. Let $\{N_t(\cdot), t \in \mathbb{Z}\}$ be a sequence of independent Poisson

processes with unit intensity. Then Y_t in (3.2.1) can be restated in terms of $N_t(\cdot)$ by assuming that

$$Y_t = N_t(\lambda_t), \tag{3.2.2}$$

i.e., Y_t is equal to the number of events of $N_t(\cdot)$ in the time interval $[0, \lambda_t]$.

The main difficulty of investigating the stability properties of $\{\lambda_t\}$ comes from the fact that the state space of the conditional distribution of λ_t given $\mathcal{F}_{t-1}$ depends on Y_{t-1}, which only takes discrete values. In particular, it is easy to show that $\{\lambda_t\}$ is not a strong Feller chain even for the INGARCH model without a threshold. To see this, consider $\lambda_t = d + a\lambda_{t-1} + bY_{t-1}$ for $t \geq 1$ with some initial value λ_0. Assume that d, a and b are all rational, and $f = \mathbf{1}_{\mathbb{Q}}$. Then $Pf = f$ is not continuous, hence the chain is not strong Feller. This means that one needs to apply more nonstandard Markov chain theory, such as Lyapunov's method and e-chains, in order to establish stability properties. Readers can refer to Sections 6.1-6.2 in Duflo (1997), Definition 6.2.10 and Theorem 6.2.2 for the corresponding definitions and theory. We begin with the following proposition establishing the stationarity of $\{\lambda_t\}$.

Proposition 3.2.1. *Consider model (3.2.1) and assume that $a_2 + b_2 < 1$. Then the Markov chain $\{\lambda_t\}$ has at least one invariant probability measure (i.p.m). In addition, if $a_1 < 1$, then the stationary distribution, denoted by π, is unique.*

Proof. We first show the existence of a stationary distribution provided $a_2 + b_2 < 1$. For any $x \in E$, which is the state space of $\{\lambda_t\}$, denoting $\bar{d} = \max\{d_1, d_2\}$, $d = d_1 - d_2$, $a = a_1 - a_2$ and $b = b_1 - b_2$, we have

$$\begin{aligned} \mathrm{E}(\lambda_2|\lambda_1 = x) &= \mathrm{E}\{d_2 + a_2\lambda_1 + b_2 Y_1 + (d + a\lambda_1 + bY_1)\mathbf{1}_{[Y_1 \leq r]}|\lambda_1 = x\} \\ &\leq \bar{d} + (a_2 + b_2)x + br + axP(Y_1 \leq r|\lambda_1 = x), \end{aligned}$$

where $Y_1 \sim \text{Pois}(x)$. Note that $\text{E}(Y_1|1 \leq Y_1 \leq r+1) = \sum_{k=1}^{r+1} kP(Y_1 = k|1 \leq Y_1 \leq r+1) = \{\sum_{k=1}^{r+1} kx^k/k!e^{-x}\}/P(1 \leq Y_1 \leq r+1) = \{x\sum_{k=0}^{r} x^k/k!e^{-x}\}/P(1 \leq Y_1 \leq r+1) = xP(Y_1 \leq r)/P(1 \leq Y_1 \leq r+1)$, so it implies that $xP(Y_1 \leq r) \leq \text{E}(Y_1|1 \leq Y_1 \leq r+1) \leq r+1$. Hence we have

$$\text{E}(\lambda_2|\lambda_1 = x) \leq \bar{d} + a(r+1) + br + (a_2 + b_2)x.$$

It follows that

$$\begin{aligned}
\text{E}(\lambda_3|\lambda_1 = x) &= \text{E}[\text{E}(\lambda_3|\lambda_2)|\lambda_1 = x] \\
&\leq \text{E}[\bar{d} + a(r+1) + br + (a_2 + b_2)\lambda_2|\lambda_1 = x] \\
&\leq K_1(1 + a_2 + b_2) + (a_2 + b_2)^2 x,
\end{aligned}$$

where $K_1 = \bar{d} + a(r+1) + br$. By induction, we have for any $t \geq 1$,

$$\text{E}(\lambda_t|\lambda_1 = x) \leq K_1[1 + (a_2 + b_2) + \ldots + (a_2 + b_2)^{t-2}] + (a_2 + b_2)^{t-1}x.$$

For any $\epsilon > 0$, define the compact subset $C = [0, K]$, where $K \geq \{K_1/(1 - a_2 - b_2) + x\}/\epsilon$. Then the t-th step transition probability is

$$\begin{aligned}
P^t(x, C) &= P(\lambda_{t+1} \leq K|\lambda_1 = x) \geq 1 - \text{E}(\lambda_{t+1}|\lambda_1 = x)/K \\
&\geq 1 - \{K_1\frac{1 - (a_2 + b_2)^t}{1 - (a_2 + b_2)} + (a_2 + b_2)^t x\}/K \\
&\geq 1 - \{K_1/(1 - (a_2 + b_2)) + x\}/K \geq 1 - \epsilon.
\end{aligned}$$

Hence $\{1/k\sum_{t=1}^{k} P^t(x, \cdot), k \geq 1\}$ is tight, i.e., $\{\lambda_t\}$ is bounded in probability on average. Since $\{\lambda_t\}$ is a weak Feller chain (see Wang *et al.* (2012)), it follows that $\{\lambda_t\}$ has at least one invariant probability measure (see Meyn and Tweedie (2009)).

In what follows, we further assume $a_1 < 1$ and prove the uniqueness of the stationary distribution. First note that $\lambda^* = d_1/(1 - a_1)$ is a reachable state by letting

$Y_1 = Y_2 = \ldots = Y_t = 0$ for large t. What remains is to show that $\{\lambda_t\}$ is an e-chain, i.e., for any continuous function f with compact support and $\epsilon > 0$, there exists an $\eta > 0$ such that $|P^k_{x_1} f - P^k_{z_1} f| < \epsilon$, for $|x_1 - z_1| < \eta$ and all $k \geq 1$, where $P^k_{x_1} f = \mathrm{E}\{f(\lambda_k)|\lambda_0 = x\}$ (see Definition 6.2.10). Without loss of generality, assume $|f| \leq 1$. Take ϵ' and η sufficiently small such that $\epsilon' + 4\eta/(1-\bar{a}) < \epsilon$, where $\bar{a} = \max\{a_1, a_2\} < 1$, and $|f(x_1) - f(z_1)| < \epsilon'$ whenever $|x_1 - z_1| < \eta$. Denote $p(\cdot|x)$ as the probability mass function of a Poisson distribution with intensity x. Then for the case $k = 1$,

$$\begin{aligned} |P_{x_1} f - P_{z_1} f| &\leq |\sum_{i=0}^{r} f(d_1 + a_1 x_1 + b_1 i) p(i|x_1) - \sum_{i=0}^{r} f(d_1 + a_1 z_1 + b_1 i) p(i|z_1)| \\ &\quad + |\sum_{j=r+1}^{\infty} f(d_2 + a_2 x_1 + b_2 j) p(j|x_1) - \sum_{j=r+1}^{\infty} f(d_2 + a_2 z_1 + b_2 j) p(j|z_1)| \\ &:= I + II. \end{aligned}$$

Note that for $x_1 \geq z_1$,

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

and when $x_1 < z_1$, we also have

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

Hence for any x_1 and z_1, we have

$$\sum_{i=0}^{\infty} |p(i|x_1) - p(i|z_1)| \leq 2(1 - e^{-|x_1 - z_1|}). \tag{3.2.3}$$

So it follows that

$$\begin{aligned} I &\leq \sum_{i=0}^{r} |f(d_1 + a_1 x_1 + b_1 i) - f(d_1 + a_1 z_1 + b_1 i)| p(i|x_1) \\ &\quad + \sum_{i=0}^{r} |f(d_1 + a_1 z_1 + b_1 i)||p(i|x_1) - p(i|z_1)| \\ &\leq \epsilon' F(r|x_1) + 2(1 - e^{-|x_1 - z_1|}), \end{aligned}$$

where $F(r|x_1) = \sum_{i=0}^{r} p(i|x_1)$. The last inequality follows from (3.2.3), $|f| \leq 1$ and the fact that $|(d_1 + a_1 x_1 + b_1 i) - (d_1 + a_1 z_1 + b_1 i)| = a_1 |x_1 - z_1| < \eta$. It follows from a similar argument that $II \leq \epsilon'(1 - F(r|x_1)) + 2(1 - e^{-|x_1 - z_1|})$. Hence we have

$$|P_{x_1} f - P_{z_1} f| \leq \epsilon' + 4(1 - e^{-|x_1 - z_1|}) \tag{3.2.4}$$

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

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

that

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

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

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

which follows from (3.2.3) and (3.2.4). Similarly, we have $IV \leq [\epsilon' + 4(1 - e^{-|x_2^{(2)} - z_2^{(2)}|})](1 - F(r|x_1)) + 2(1 - e^{-|x_1 - z_1|})$. Since $|x_2^{(1)} - z_2^{(1)}| = a_1|x_1 - z_1|$ and $|x_2^{(2)} - z_2^{(2)}| = a_2|x_1 - z_1|$, so by letting $\bar{a} = \max\{a_1, a_2\}$, we have

$$|P_{x_1}^2 f - P_{z_1}^2 f| \leq \epsilon' + 4(1 - e^{-\bar{a}|x_1 - z_1|}) + 4(1 - e^{-|x_1 - z_1|}).$$

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

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

where the second inequality holds since $1 - e^{-x} \leq x$. Hence $\{\lambda_t\}$ is an e-chain. Together with the existence of one reachable state, it is shown that the stationary distribution is unique, which completes the proof. □

Hence according to Proposition 3.2.1, it is assumed that $a_2 + b_2 < 1$ and $a_1 < 1$ for the rest of the discussion to guarantee the uniqueness of the stationary distribution. Furthermore, a proposition concerning the strong law of large numbers on the $\{\lambda_t\}$ process can be derived.

Proposition 3.2.2. *Consider model (3.2.1) and assume that $a_2 + b_2 < 1$ and $a_1 < 1$. Then*

1. *The Markov chain* $\{\lambda_t\}$ *has finite moments of all orders.*

2. *For any* μ*-a.s. continuous function* ϕ *satisfying*

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

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

$$\frac{1}{n}[\phi(\lambda_1) + \ldots + \phi(\lambda_n)] \longrightarrow \mu(\phi), \; a.s.$$

for any initial value $\lambda_0 = x$*.*

The proof of Proposition 3.2.2 relies on the following lemmas. The first one deals with moments of a Poisson random variable and uses Stirling numbers of the second kind, denoted by $\left\{ {n \atop k} \right\}$. For $n \geq 0$ and $0 \leq k \leq n$, they satisfy the recurrence:

$$\left\{ {n \atop k} \right\} = \left\{ {n-1 \atop k-1} \right\} + k \left\{ {n-1 \atop k} \right\},$$

with $\left\{ {n \atop n} \right\} = 1$ for $n \in \{0, 1, 2, \ldots\}$, $\left\{ {n \atop 0} \right\} = 0$ for $n \in \mathbb{N} = \{1, 2, \ldots\}$ and $\left\{ {n \atop k} \right\} = 0$ if $k > n$ (see Ferland *et al.* (2006)).

Lemma 3.2.1. *Let* X *be a Poisson random variable with mean* λ*. Then the moments of* X *satisfy*

$$E[X^k] = \sum_{i=0}^{k} \left\{ {k \atop i} \right\} \lambda^i.$$

The detailed proof can be found in Hardy (1996).

Lemma 3.2.2. *For a Poisson process* $\{N(u), u \geq 0\}$ *with a unit rate, we have*

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

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

Proof. The first result apparently holds for $u \in \mathbb{Z}$ according to the law of large numbers. As for an arbitrary $u \in \mathbb{R}_+$, let $\lfloor u \rfloor$ be the integer part of u, then it follows from $\lfloor u \rfloor \leq u < \lfloor u \rfloor + 1$ that

$$N(\lfloor u \rfloor) \leq N(u) \leq N(\lfloor u \rfloor + 1).$$

Therefore $\lim_{u\to\infty} N(u)/u = 1$.

To prove the uniform integrability of $\{[N(u)/u]^k, u \geq 1\}$, note that for any integer-valued $s \geq k$, it follows from Lemma 3.2.1 that $\mathrm{E}[N(u)/u]^s$ is bounded for all $u \geq 1$. So the family $\{(N(u)/u)^k, u \geq 1\}$ is uniformly integrable. □

Lemma 3.2.3. *For $s \in \mathbb{N} = \{1, 2, \ldots\}$, let $V(\lambda) = \lambda^s$. Then*

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

Proof. First note that

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

For ω fixed, it follows from Lemma 3.2.2 that $Y_0/\lambda \to 1$ and $\mathbf{1}_{[Y_0\leq r]} \to 0$ as $\lambda \to \infty$. Hence $h(\lambda, \omega) \to (a_2 + b_2)^s$ almost surely as $\lambda \to \infty$.

It remains to show that $\{h(\lambda, \omega), \lambda \geq 1\}$ is uniformly integrable. It follows from the inequality $(a + b)^s \leq 2^{s-1}(a^s + b^s)$ that there exists a constant C such that

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

for $\lambda \geq 1$. Since $\{Y_0/\lambda, \lambda \geq 1\}$ is uniformly integrable according to Lemma 3.2.2, so is the family $\{h(\lambda, \omega), \lambda \geq 1\}$. Hence the limit and integration are exchangeable, which completes the proof. □

Proof of Proposition 3.2.2. The first result is a direct application of Lemma 3.2.3. The strong law of large numbers also follows from this method, see Proposition 6.2.12 and the remarks after it in Duflo (1997). □

The law of the large numbers serves an important role in establishing the asymptotic theory of the parameter estimates when transitioning to the estimation stage in the next section. Note that the properties of the count process $\{Y_t\}$ can be easily deduced from those of $\{\lambda_t\}$, and are recorded in the following corollary.

Corollary 3.2.1. *Under the same conditions of Proposition 3.2.2, the bivariate process* $\{(\lambda_t, Y_t)\}$ *has a unique stationary distribution and* $\{Y_t\}$ *has finite moments of all orders.*

3.2.2 Likelihood Inference

We consider the maximum likelihood estimates of the parameters in model (3.2.1). For a fixed threshold parameter r, one can obtain the estimates of $(d_1, a_1, b_1, d_2, a_2, b_2)$ by maximizing the log likelihood function conditional on the initial value λ_1. Then r is estimated by maximizing the likelihood based on a grid search, which is over the integer numbers in the interval $[0, r^*]$. r^* should be large enough to guarantee the coverage of the true value of r, but meanwhile one should make sure that there are sufficient observations in each regime to obtain quality estimates of the parameters. In practice, one can choose the range to be $[\hat{q}_1, \hat{q}_2]$, where $\hat{q}_1$ and $\hat{q}_2$ are sample quantiles.

To investigate the asymptotic behavior of the MLE, recall that $\theta^{(i)} = (d_i, a_i, b_i)^T$ is the parameter vector for the i^{th}-regime, where $i = 1, 2$. Then denote $\theta = (r, \theta^{(1)T}, \theta^{(2)T})^T$ as the parameter vector of model (3.2.1) and θ_0 as the true values. Let $\lambda_{t,i} = d_i + a_i\lambda_{t-1} + b_iY_{t-1}$ for $i = 1, 2$, then $\lambda_t = \sum_i \lambda_{t,i}\mathbf{1}_{[Y_t \in R_i]}$. Let $\{\tilde{\lambda}_t\}_{t=1}^n$ be the sequence generated by (3.2.1) with an initial value $\tilde{\lambda}_1$, then the log likelihood function,

up to a constant free of θ, is given by

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

where $\tilde{\ell}_t = -\tilde{\lambda}_t + Y_t \log(\tilde{\lambda}_t)$. Then the MLE of θ is given by

$$\hat{\theta}_n = \text{argmax}_{\theta \in [0, r^*] \times \mathcal{D}} \tilde{\ell}(\theta),$$

where r^* is a large positive integer and $\mathcal{D}$ is a compact subset of $\mathcal{R}^6$ that will be defined later. It can be shown that the asymptotic behavior of the parameter estimates does not depend on the choice of $\tilde{\lambda}_1$, and relies on the following assumptions on the underlying process and the parameter space.

(A1) The observations $\{Y_t\}_{t=1}^n$ are generated from a self-excited threshold Poisson autoregression process, with the true parameter vector $\theta_0 \in [0, r_*] \times \mathbf{\Theta}$ and $\mathbf{\Theta} = \{(d_1, a_1, b_1, d_2, a_2, b_2)^T \in \mathbb{R}_+^6 : a_1 < 1, b_1 < 1, a_2 + b_2 < 1\}$, where $\mathbb{R}_+ = (0, \infty)$.

(A2) The estimation is searched over $\mathcal{D}$, where $\mathcal{D}$ is a compact subset of $\mathbf{\Theta}$ and $\theta_0 \in \mathbb{N} \times \mathcal{D}^o$.

Note that although it is shown in Proposition 3.2.1 that $\{(\lambda_t, Y_t)\}$ has a unique stationary distribution without the condition that $b_1 < 1$, it is much more convenient to assume so when proving the asymptotic properties of the MLE. We conjecture that the same asymptotic properties would hold in absence of this constraint and leave it for future investigation.

Firstly, the strong consistency of $\hat{\theta}_n$ is established under the assumptions formulated above.

Theorem 3.2.1. *Consider model (3.2.1) and assume that (A1) and (A2) hold. Then the MLE $\hat{\theta}_n$ of the parameter vector θ is strongly consistent, i.e.,*

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

Since $r \in \mathbb{Z}$, Theorem 3.2.1 implies that $\hat{r} = r$ when the sample size is sufficiently large. We henceforth remove r from the parameter vector θ and only consider a central limit theorem for the maximum likelihood estimator with known threshold r. Therefore, $\tilde{\ell}$ is differentiable with respect to θ, and the score function can be calculated recursively, which is given by

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

where $\partial \tilde{\lambda}_t / \partial \theta = (\partial \tilde{\lambda}_t / \partial \theta^{(1)T}, \partial \tilde{\lambda}_t / \partial \theta^{(2)T})^T$ and $\partial \tilde{\lambda}_t / \partial \theta^{(i)} = (1, \tilde{\lambda}_{t-1}, Y_{t-1})^T \mathbf{1}_{[Y_{t-1} \in R_i]} + a_{t-1} \partial \tilde{\lambda}_{t-1} / \partial \theta^{(i)}$ for $i = 1, 2$. Now define

$$G = \mathrm{E}\left[\frac{1}{\lambda_t} \left(\frac{\partial \lambda_t}{\partial \theta} \right) \left(\frac{\partial \lambda_t}{\partial \theta} \right)^T \right], \tag{3.2.5}$$

then the asymptotic distribution of the maximum likelihood estimator is demonstrated in the following theorem.

Theorem 3.2.2. *Assume model (3.2.1) and that assumptions (A1) and (A2) hold. Then the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically normal, i.e.,*

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

Furthermore, the matrix G in (3.2.5) can be estimated consistently by

$$\widehat{G} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\tilde{\lambda}_t} \left(\frac{\partial \tilde{\lambda}_t}{\partial \theta} \right) \left(\frac{\partial \tilde{\lambda}_t}{\partial \theta} \right)^T. \tag{3.2.6}$$

The proofs of Theorems 3.2.1 and 3.2.2 can be found in Wang *et al.* (2012).

3.2.3 Simulation

The performance of the estimation is examined through a simulation study. Two sets of parameters are considered, where the first set has $a_i + b_i < 1$ for $i = 1, 2$

and the second set has $a_1 + b_1 > 1$ and yields negative autocorrelation. Each sample size and parameter configuration is replicated 1000 times, and the results reported are averaged over these 1000 replications. According to the estimation procedure described above, we consider the search range for r to be between the 0.2 and 0.8 empirical quantiles of the data.

According to Theorems 3.2.1 and 3.2.2, the sample means and variances are expected to be close to the theoretical values as sample size grows large enough. For the threshold parameter r, its estimate $\hat{r}$ is supposed to be identical to the true value. The simulation results for two sets of parameters are reported in Table 3.1 and Table 3.2, respectively. The standard errors of the MLE $\hat{\theta}_n$ from all the replications are calculated, and are compared to the asymptotic standard deviations of the MLE, which are computed as the square root of the diagonal elements of $\hat{G}^{-1}$ obtained in (3.2.6).

It can be observed that $\hat{r}$ converges to r very fast. However the speed of this convergence seems to be dependent on other parameters. For the first set of parameters, even when n is as large as 3000, $\hat{r}$ does not equal to r in all samples. However, $\hat{r}$ is identical to the true value even when the sample size is 500 for the second set of parameters, which is a moderate sample size for the threshold model. The consistency and asymptotic variance of the other parameters are confirmed in both examples. The average estimated parameters are close to the true values, the accuracy increases as the sample size increases. However, the intercept parameters d_i seem to have large variances, comparing to the other parameters. The large variance of the intercept was also found in estimating an INGARCH model in Fokianos *et al.* (2009) and Davis and Liu (2012). In the first example, s.e.$(\hat{\theta}_n)$ and a.s.d$(\hat{\theta}_n)$ match each other reasonably well. However, such phenomenon is not so apparent in the second example, especially for d_1 and d_2. This might be due to the fact that the lower regime is explosive in the

Table 3.1: Results of Simulation 1 for the SETINGARCH model: s.e.($\hat{\theta}_n$) is the standard error of MLE from 1000 replications, a.s.d.($\hat{\theta}_n$) is the averaged asymptotic deviation from 1000 replications calculated according to (3.2.6).

Sample size	Description	r	d_1	a_1	b_1	d_2	a_2	b_2
	θ_0	7	0.50	0.70	0.20	0.30	0.40	0.50
$n = 500$	$\overline{\hat{\theta}_n}$	6.80	0.63	0.69	0.18	0.83	0.37	0.47
	s.e.($\hat{\theta}_n$)	1.05	0.23	0.05	0.04	0.65	0.09	0.08
	a.s.d($\hat{\theta}_n$)	N/A	0.20	0.045	0.05	0.67	0.07	0.08
$n = 1000$	$\overline{\hat{\theta}}$	7.00	0.56	0.70	0.19	0.60	0.38	0.48
	s.e.($\hat{\theta}_n$)	0.71	0.19	0.04	0.05	0.66	0.08	0.08
	a.s.d($\hat{\theta}_n$)	N/A	0.17	0.04	0.04	0.64	0.07	0.07
$n = 2000$	$\overline{\hat{\theta}}$	7.02	0.53	0.70	0.20	0.42	0.39	0.49
	s.e.($\hat{\theta}_n$)	0.35	0.16	0.04	0.04	0.54	0.07	0.07
	a.s.d($\hat{\theta}_n$)	N/A	0.16	0.04	0.04	0.59	0.07	0.07
$n = 3000$	$\overline{\hat{\theta}}$	7.00	0.52	0.70	0.20	0.37	0.40	0.50
	s.e.($\hat{\theta}_n$)	0.07	0.16	0.04	0.04	0.52	0.07	0.07
	a.s.d($\hat{\theta}_n$)	N/A	0.16	0.04	0.04	0.58	0.07	0.07

Table 3.2: Results of Simulation 2 for the SETINGARCH model: s.e.($\hat{\theta}_n$) is the standard error of MLE from 1000 replications, a.s.d.($\hat{\theta}_n$) is the averaged asymptotic deviation from 1000 replications calculated according to (3.2.6).

Sample size	Description	r	d_1	a_1	b_1	d_2	a_2	b_2
	θ_0	6	0.50	0.80	0.70	0.20	0.20	0.10
$n = 500$	$\overline{\hat{\theta}_n}$	6.00	0.47	0.82	0.69	0.32	0.19	0.09
	s.e.($\hat{\theta}_n$)	0	0.17	0.06	0.05	0.26	0.04	0.04
	a.s.d($\hat{\theta}_n$)	N/A	0.19	0.06	0.05	0.37	0.04	0.04
$n = 1000$	$\overline{\hat{\theta}_n}$	6.00	0.50	0.81	0.70	0.28	0.20	0.09
	s.e.($\hat{\theta}_n$)	0	0.17	0.06	0.05	0.28	0.04	0.04
	a.s.d($\hat{\theta}_n$)	N/A	0.18	0.06	0.05	0.37	0.04	0.04
$n = 2000$	$\overline{\hat{\theta}_n}$	6.00	0.50	0.80	0.70	0.23	0.20	0.10
	s.e.($\hat{\theta}_n$)	0	0.17	0.06	0.05	0.29	0.04	0.04
	a.s.d($\hat{\theta}_n$)	N/A	0.18	0.06	0.05	0.37	0.04	0.04
$n = 3000$	$\overline{\hat{\theta}_n}$	6.00	0.50	0.80	0.70	0.22	0.20	0.10
	s.e.($\hat{\theta}_n$)	0	0.18	0.06	0.05	0.32	0.04	0.04
	a.s.d($\hat{\theta}_n$)	N/A	0.18	0.06	0.05	0.37	0.04	0.04

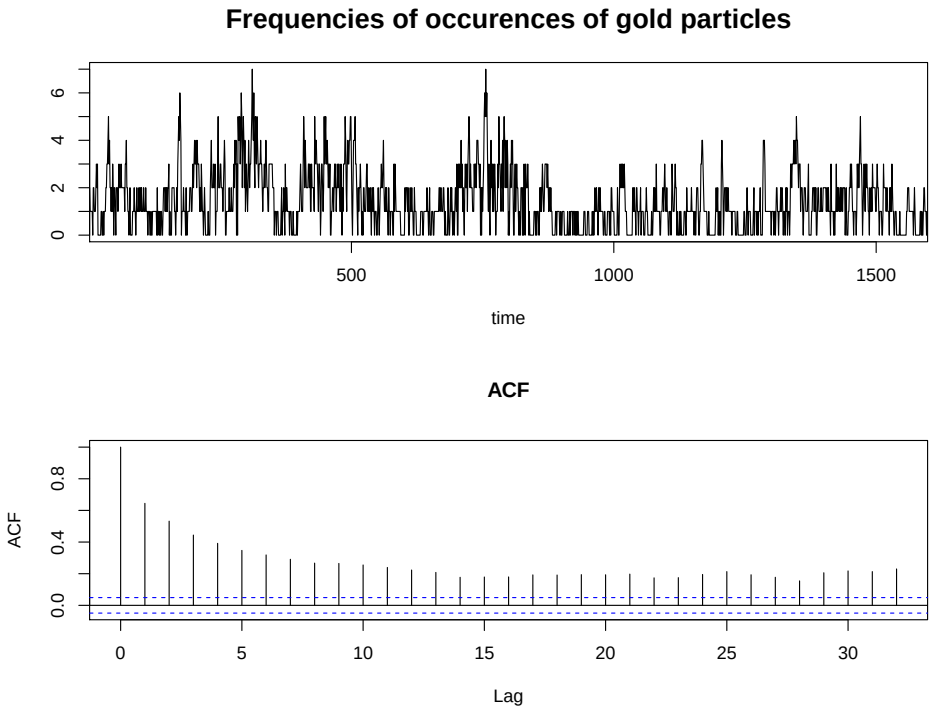

Figure 3.1: Frequencies of occurrences of gold particles and the ACF plot.

second example.

3.2.4 Real Data Application

Gold particles are observed at constant intervals of time which are never less than a few hundredths of a second. Then the frequency of occurences is written down and a time series of counts is generated. This data file is a part of the data set printed in Guttorp (1991), and is exhibited in Figure 3.1 together with the plot of the autocorrelation.

For comparison, both of the Poisson INGARCH (2.4.6) and the self-excited thresh-

Table 3.3: Summary of model estimates for the gold particles.

	INGARCH	SETINGARCH	SETINGARCH with $d_2 = 0$
d_1	0.1915 (0.030)	0.1557 (0.034)	0.1540 (0.034)
a_1	0.4345 (0.041)	0.5366 (0.049)	0.5399 (0.049)
b_1	0.4308 (0.030)	0.3030 (0.048)	0.3026 (0.049)
d_2		0.0673 (0.157)	
a_2		0.3036 (0.069)	0.3065 (0.068)
b_2		0.5781 (0.078)	0.6009 (0.078)
r		1	1
LogL	-2082.26	-2076.64	-2076.64
AIC	4170.52	4167.29	4165.47

old Poisson autoregression (3.2.1) models are fitted to this data set, and the estimation results are summarized in Table 3.3, where the numbers in the parentheses are the standard errors of the estimates. The original fitting of the SETINGARCH model gives an insignificant estimate of d_2, so another model restricting $d_2 = 0$ is also provided. Note that $d_2 = 0$ is allowed in the formulation of model (3.2.1). It turns out that the SETINGARCH has a greater log likelihood function and reduces the AIC. The prediction scores (see (2.5.1)-(2.5.3)) are very close to those of the Poisson INGARCH model. The fitted conditional mean processes of the SETINGARCH with $d_2 = 0$ are depicted in Figure 3.2, where the dashed horizontal line represents the level of the threshold. It can be seen that $\hat{\lambda}_t$ is capable of capturing the fluctuation of the observations. The fitted conditional mean process using the Poisson INGARCH model lies very close to the curve in the plot, so for the clarity of the graph, it is not shown in the figure. The plot of the Pearson residuals is also given in Figure 3.2, which suggests that the residuals appear to be white.

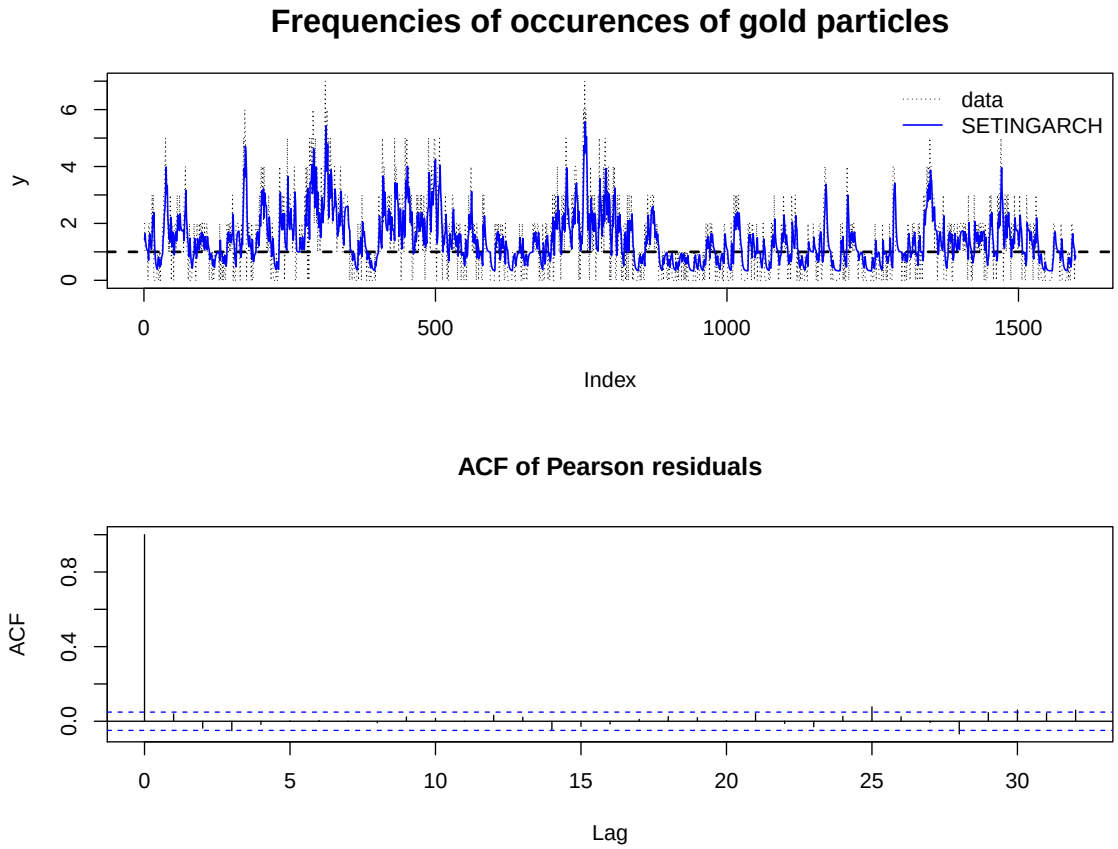

Figure 3.2: Top: the black dashed curve represents the observations, the blue solid one describes the fitted conditional mean process using SETINGARCH and the dashed horizontal line is the threshold; bottom: ACF of standardized Pearson residuals.

3.3 INGARCH with Markovian Covariates

In this section, we consider the problem of entering covariates into an INGARCH model based on a Poisson or negative binomial distribution. In particular, we assume that the covariates are a realization of some Markov process.

3.3.1 Model Formulation and Stability Properties

Define $\mathcal{F}_0^{Y,Z} = \sigma\{\mu_1, Z_1\}$, which is assumed fixed for the moment. Let $Y_t \in \mathbb{N}_0$ be the observation at time t, and $Z_t = (Z_{t,1}, \ldots, Z_{t,p})^T$ be the p-dimensional vector containing the covariates information at time t. It is assumed throughout that $\{Z_t\}$ is a time-homogeneous Markov chain. A *Poisson INGARCH with covariates* is defined as, for all $t \geq 1$,

$$\mathcal{L}(Z_t|\mathcal{F}_{t-1}^{Y,Z}) = \mathcal{L}(Z_t|Z_{t-1}) \text{ and } Y_t|\mathcal{F}_{t-1}^{Y,Z}, Z_t \sim \text{Pois}(e^{Z_t^T\gamma}\mu_t),$$

where

$$\mu_t = \delta + \alpha\mu_{t-1} + \beta Y_{t-1}e^{-Z_{t-1}^T\gamma}. \tag{3.3.1}$$

Here $\mathcal{F}_t^{Y,Z} = \sigma\{\mu_1, Y_1, \ldots, Y_t, Z_1, \ldots, Z_t\}$, $\mathcal{L}(X)$ represents the distribution of a random variable X, and $\delta > 0, \alpha, \beta \geq 0, \gamma$ are parameters. A *negative binomial INGARCH with covariates* can be defined similarly as

$$\mathcal{L}(Z_t|\mathcal{F}_{t-1}^{Y,Z}) = \mathcal{L}(Z_t|Z_{t-1}) \text{ and } Y_t|\mathcal{F}_{t-1}^{Y,Z}, Z_t \sim \text{NB}(r, p_t),$$

where $r(1-p_t)/p_t = e^{Z_t^T\gamma}\mu_t$ and

$$\mu_t = \delta + \alpha\mu_{t-1} + \beta Y_{t-1}e^{-Z_{t-1}^T\gamma}. \tag{3.3.2}$$

Note that by recursion, we have for all $l \geq 1$,

$$\mu_t = \delta(1-\alpha^{l+1})/(1-\alpha) + \beta\sum_{k=0}^{l}\alpha^k Y_{t-k-1}e^{-Z_{t-k-1}^T\gamma} + \alpha^{l+1}\mu_{t-l-1}. \tag{3.3.3}$$

It then follows from $Y_t \geq 0$ that $\mu_t \geq \mu^* := \delta/(1-\alpha)$ for all t provided that $\alpha < 1$. Define $X_t = (\mu_t, Z_t^T)^T$, then it is easy to see that $\{X_t\}$ is a time-homogeneous Markov chain with state space $E = [\mu^*, +\infty) \times H$, where H is the state space of $\{Z_t\}$. The conditions under which $\{X_t\}$ enjoys stability properties will be formulated in the following proposition.

Proposition 3.3.1. *Consider the Poisson model (3.3.1) or the negative binomial model (3.3.2) and assume that $\alpha + \beta < 1$. Further assume that $\{Z_t\}$ satisfies the following conditions:*

(Z1) $\{Z_t\}$ is weak Feller with a compact state space $H \subset \mathbb{R}^p$ and a unique stationary distribution, which implies that $e^{Z_t^T \gamma} \in [1/M, M]$ for some $M > 0$ for all t.

(Z2) $\{Z_t\}$ is φ-irreducible for some measure φ defined on H. In particular, for any $z \in H$ and $C \subset H$ with $\varphi(C) > 0$, the transition probability $\mathbf{Q}(z, C) > 0$.

(Z3) For any $z \in H$ and Borel set $B \subset [1/M, M]$ with $\lambda(B) > 0$, where $\lambda(\cdot)$ is the Lebesgue measure, $P(e^{Z_2^T \gamma} \in B | Z_1 = z) > 0$.

Then the Markov process $X_t = (\mu_t, Z_t^T)^T$ is geometrically ergodic, hence is stationary and ergodic when initiated from its stationary distribution.

Remark 3.3.1. Although the assumptions on $\{Z_t\}$ in Proposition 3.3.1 are due to technical reasons, (Z3) can be easily justified in practice if at least one of the covariate variables takes a wide range of continuous values. However, in the presence of discrete regressors, (Z2) has its own limitation, since the choice of φ could be difficult.

Proof. First note that μ^* is reachable from any initial value of X_1, i.e., for any $X_1 = (\mu_1, Z_1^T)^T$ and $\epsilon > 0$, there exits $t \geq 1$ such that $P(|\mu_t - \mu^*| < \epsilon | \mu_1, Z_1) > 0$. This

follows from (3.3.3) by letting t be large enough and $Y_{t-1} = \ldots = Y_1 = 0$. For any $s > \mu^*$, let $Y_1 = N$, and $Y_2 = \ldots = Y_{t-1} = 0$, then according to (3.3.3) we have

$$\mu_t(N) = \frac{\delta(1-\alpha^{t-1})}{1-\alpha} + \alpha^{t-2}(\beta N e^{-Z_1^T\gamma} + \alpha\mu_1).$$

Note that $\mu_t(N) - \mu_t(N-1) = \alpha^{t-2}\beta e^{-Z_1^T\gamma}$, so for any $\epsilon > 0$, there exists t large enough such that $\mu_t(N) - \mu_t(N-1) < \epsilon$ and $\mu_t(0) = \delta(1-\alpha^{t-1})/(1-\alpha) + \alpha^{t-1}\mu_1 < \mu^* + \alpha^{t-1}\mu_1 < s$. For such t, pick $N \in \mathbb{N}$ so that $\mu_t(N-1) < s$ and $\mu_t(N) \geq s$. Then $|\mu_t(N) - s| < \epsilon$, which implies that $P(|\mu_t - s| < \epsilon | \mu_1, Z_1) > 0$.

Now we show that $\{X_t\}$ is ϕ-irreducible, where $\phi = \varphi_0 \times \varphi$ and φ_0 is the Lebesgue measure defined on $[\mu^* + \beta M, \infty)$. Consider $X_1 = (\mu_1, z_1^T)^T \in E$, and $A = (A_1, A_2) \subset E$ with $\phi(A) > 0$. Without loss of generality, assume that $A_1 \subset [\mu^* + \beta M, \infty)$. Since $\varphi_0(A_1) = \sup\{\varphi_0(K) : K \subset A_1 \text{ and } K \text{ is compact}\}$, it follows that there exists a compact set $K \subset A_1$ with $\varphi_0(K) > 0$. We claim that for this K, there exists $c \in K$ such that for any $\epsilon > 0$, $\varphi_0(K \cap (c-\epsilon, c+\epsilon)) > 0$. To see this, assume that for all $c \in K$, there exists an $\epsilon(c) > 0$, such that $\varphi_0(K \cap (c - \epsilon(c), c + \epsilon(c))) = 0$. Then one can find a finite open covering $\{\mathcal{O}_i\}_{i=1}^m$ such that $\varphi_0(K \cap \bigcup_{i=1}^m \mathcal{O}_i) = 0$. This implies that $\varphi_0(K) \leq \sum_{i=1}^m \varphi_0(K \cap \mathcal{O}_i) = 0$, which contradicts the fact that $\varphi_0(K) > 0$. Since any $s \geq \mu^*$ is reachable, so there exists $t \geq 1$ such that $\mu^* \leq (c - \delta + \beta/M)/\alpha < \mu_t < (c - \delta + \beta M)/\alpha$. It follows that there exists $\epsilon > 0$ such that $\delta + \alpha\mu_t + \beta/M < c - \epsilon < c < c + \epsilon < \delta + \alpha\mu_t + \beta M$. Since $B := K \cap (c-\epsilon, c+\epsilon) \subset A_1$, we have

$$\begin{aligned}
P(\mu_{t+1} \in A_1 | \mu_t, \mu_1, Z_1) &\geq P(\mu_{t+1} \in B | \mu_t, \mu_1, Z_1) \\
&= P(\delta + \alpha\mu_t + \beta Y_t e^{-Z_t^T\gamma} \in B | \mu_t, \mu_1, Z_1) \\
&\geq P(\delta + \alpha\mu_t + \beta e^{-Z_t^T\gamma} \in B | \mu_t, Y_t = 1, \mu_1, Z_1) P(Y_t = 1 | \mu_t, \mu_1, Z_1).
\end{aligned}$$

Since $\delta + \alpha\mu_t + \beta e^{-Z_t^T\gamma}$ has positive density everywhere on $(\delta + \alpha\mu_t + \beta/M, \delta + \alpha\mu_t + \beta M)$, and $B \subset (c-\epsilon, c+\epsilon) \subset (\delta + \alpha\mu_t + \beta/M, \delta + \alpha\mu_t + \beta M)$, so it follows from assumptions

(Z2) and (Z3) that $P(\mu_{t+1} \in A_1|\mu_t, \mu_1, Z_1) > 0$. Hence there exists $t \geq 1$ such that $P(\mu_{t+1} \in A_1|\mu_1, z_1) > 0$. Moreover, $P(\mu_{t+1} \in A_1, Z_{t+1} \in A_2|\mu_1, Z_1) = P(\mu_{t+1} \in A_1|\mu_1, Z_1)P(Z_{t+1} \in A_2|\mu_{t+1} \in A_1, \mu_1, Z_1) > 0$ according to the assumptions on $\{Z_t\}$. Therefore $\{X_t\}$ is ϕ-irreducible.

Define the drift function $V : E \to \mathbb{R}$ as $V(x) = 1+x_1$, where $x = (x_1, x_2, \ldots, x_{p+1})^T$. Since $\alpha + \beta < 1$, there exists a $\rho > 0$ such that $\alpha + \beta + \rho < 1$, and let $D = [\delta/(1-\alpha), (\delta+\rho)/(1-\alpha-\beta-\rho)] \times H$. Then for $x_1 = (\mu_1, z_1^T)^T \notin D$, i.e., $\mu_1 \geq (\delta+\rho)/(1-\alpha-\beta-\rho)$,

$$\begin{aligned} \mathrm{E}[V(X_2)|X_1 = x_1] &= \mathrm{E}(1+\mu_2|\mu_1, z_1) = 1 + \delta + \alpha\mu_1 + \beta e^{-z_1^T\gamma} e^{z_1^T\gamma}\mu_1 \\ &= 1 + \delta + (\alpha+\beta)\mu_1 \leq (1-\rho)(1+\mu_1). \end{aligned}$$

Hence it follows from Theorem 6.2.3 that $\{X_t\}$ is geometrically ergodic. All the rest results follow from the standard Markov chain theory, see Appendix. □

3.3.2 Likelihood Inference

We consider maximum likelihood estimates of the parameters in both of the Poisson-based and negative binomial-based models. Denote the $(p+3)$-dimensional parameter vector by $\theta = (\delta, \alpha, \beta, \gamma^T)^T \in (0, \infty) \times [0, \infty)^2 \times \mathbb{R}^p$. It is assumed that the distribution of $\{Z_t\}$ does not depend on θ. Then the likelihood function of model (3.3.1) conditional on μ_1 and based on the observations $Y_1, \ldots, Y_n, Z_1, \ldots, Z_n$ is given by

$$\begin{aligned} L(\theta|Y_1, \ldots, Y_n, Z_1, \ldots, Z_n, \mu_1) &= f(Y_1|\theta, Z_1, \mu_1) \prod_{t=2}^{n} f(Y_t|\mathcal{F}_{t-1}^{Y,Z}, \theta) p(Z_1) \prod_{t=2}^{n} p(Z_t|Z_{t-1}) \\ &= \prod_{t=1}^{n} \frac{(e^{Z_t^T\gamma}\mu_t)^{Y_t}}{Y_t!} e^{-\mu_t e^{Z_t^T\gamma}} p(Z_1) \prod_{t=2}^{n} p(Z_t|Z_{t-1}), \end{aligned}$$

where μ_t is updated according to the dynamics in (3.3.1). Since $p(Z_1)\prod_{t=2}^{n} p(Z_t|Z_{t-1})$ is independent of θ, so the log-likelihood function, up to a constant, is given by

$$l(\theta) = \sum_{t=1}^{n} Y_t \log \mu_t + \sum_{t=1}^{n} (Z_t^T \gamma) Y_t - \sum_{t=1}^{n} e^{Z_t^T \gamma} \mu_t. \tag{3.3.4}$$

The score function of $l(\theta)$ is

$$S_n(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n} (\frac{Y_t}{\mu_t} - e^{Z_t^T \gamma}) \frac{\partial \mu_t}{\partial \theta} + \sum_{t=1}^{n} (Y_t - \mu_t e^{Z_t^T \gamma}) \frac{\partial Z_t^T \gamma}{\partial \theta}, \tag{3.3.5}$$

where $\partial(Z_t^T\gamma)/\partial\theta = (0 \;\; 0 \;\; 0 \;\; Z_t^T)^T$, and $\partial\mu_t(\theta)/\partial\theta \in \mathbb{R}^{p+3}$ can be recursively determined by

$$\frac{\partial \mu_t(\theta)}{\partial \theta} = (1 \;\; \mu_{t-1}(\theta) \;\; Y_{t-1} e^{-Z_{t-1}^T \gamma} \;\; -\beta Y_{t-1} e^{-Z_{t-1}^T \gamma} Z_{t-1}^T)^T + \alpha \frac{\partial \mu_{t-1}(\theta)}{\partial \theta}, \tag{3.3.6}$$

and $\partial\mu_1(\theta)/\partial\theta = \mathbf{0}$. The maximum likelihood estimator $\hat{\theta}_n$ is a solution to the equation $S_n(\theta) = 0$. Furthermore, the Hessian matrix can be found by taking derivatives of the score function, i.e.,

$$\begin{aligned} H_n(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T} \;\; = \;\; & -\sum_{t=1}^{n} (\frac{Y_t}{\mu_t^2} \frac{\partial \mu_t}{\partial \theta} + e^{Z_t^T\gamma} \frac{\partial Z_t^T \gamma}{\partial \theta}) \frac{\partial \mu_t}{\partial \theta^T} + \sum_{t=1}^{n} (\frac{Y_t}{\mu_t} - e^{Z_t^T\gamma}) \frac{\partial^2 \mu_t}{\partial\theta\partial\theta^T} \\ & - \sum_{t=1}^{n} (\frac{\partial \mu_t}{\partial \theta} e^{Z_t^T\gamma} + \mu_t e^{Z_t^T\gamma} \frac{\partial Z_t^T\gamma}{\partial\theta}) \frac{\partial Z_t^T \gamma}{\partial \theta^T}, \end{aligned} \tag{3.3.7}$$

where

$$\frac{\partial^2 \mu_t}{\partial\theta\partial\theta^T} = (0 \;\; \frac{\partial \mu_{t-1}}{\partial\theta} \;\; -Y_{t-1} e^{-Z_{t-1}^T\gamma} \frac{\partial Z_{t-1}^T \gamma}{\partial\theta} \;\; \beta Y_{t-1} e^{-Z_{t-1}^T\gamma} \frac{\partial Z_{t-1}^T\gamma}{\partial\theta} Z_{t-1}^T) + \alpha \frac{\partial^2 \mu_{t-1}}{\partial\theta\partial\theta^T}, \tag{3.3.8}$$

and $\partial^2\mu_1/\partial\theta\partial\theta^T = \mathbf{0}$. Similarly, the log likelihood function of the negative binomial-based model (3.3.2), for given r and up to a constant free of θ, is given by

$$l(\theta|r) = \sum_{t=1}^{n} Y_t \log \mu_t + \sum_{t=1}^{n} (Z_t^T\gamma) Y_t - \sum_{t=1}^{n} (r + Y_t) \log(r + \mu_t e^{Z_t^T\gamma}). \tag{3.3.9}$$

The score function correspondingly is

$$S_n(\theta|r) = \sum_{t=1}^{n}\{\frac{Y_t}{\mu_t} - \frac{(r+Y_t)e^{Z_t^T\gamma}}{r+\mu_t e^{Z_t^T\gamma}}\}\frac{\partial \mu_t}{\partial \theta} + \sum_{t=1}^{n}\{Y_t - \frac{(r+Y_t)\mu_t e^{Z_t^T\gamma}}{r+\mu_t e^{Z_t^T\gamma}}\}\frac{\partial Z_t^T\gamma}{\partial \theta}. \quad (3.3.10)$$

A solution to the equation $S_n(\theta|r) = 0$ combined with a grid search on the value of r yields the maximum likelihood estimates of the parameters in model (3.3.2). In addition, the Hessian matrix can be calculated and is given by

$$\begin{aligned} H_n(\theta|r) = \frac{\partial^2 l(\theta|r)}{\partial\theta\partial\theta^T} &= \sum_{t=1}^{n}\{-\frac{Y_t}{\mu_t^2}\frac{\partial\mu_t}{\partial\theta} + (r+Y_t)\frac{e^{2Z_t^T\gamma}\frac{\partial\mu_t}{\partial\theta} - re^{Z_t^T\gamma}\frac{\partial Z_t^T\gamma}{\partial\theta}}{(r+\mu_t e^{Z_t^T\gamma})^2}\}\frac{\partial\mu_t}{\partial\theta^T} \\ &+ \sum_{t=1}^{n}\{\frac{Y_t}{\mu_t} - \frac{(r+Y_t)e^{Z_t^T\gamma}}{r+\mu_t e^{Z_t^T\gamma}}\}\frac{\partial^2\mu_t}{\partial\theta\partial\theta^T} \\ &- \sum_{t=1}^{n}\{(r+Y_t)\frac{re^{Z_t^T\gamma}\frac{\partial\mu_t}{\partial\theta} + r\mu_t e^{Z_t^T\gamma}\frac{\partial Z_t^T\gamma}{\partial\theta}}{(r+\mu_t e^{Z_t^T\gamma})^2}\}\frac{\partial Z_t^T\gamma}{\partial\theta^T}, \quad (3.3.11) \end{aligned}$$

where $\partial\mu_t/\partial\theta$ and $\partial^2\mu_t/\partial\theta\partial\theta^T$ are defined as (3.3.6) and (3.3.8), respectively.

For ease of discussion, we only investigate the asymptotic properties of the MLE of the Poisson model. However, the derivation of the asymptotic behavior of MLE of the negative binomial model is possible using similarly stylized arguments. The asymptotic behavior of the MLE of models (3.3.1) can be established based on the following assumption on the parameter space:

(AC) The true parameter vector θ_0 lies in a compact neighborhood $\Theta \in \mathbb{R}^{3+p}$ of θ_0, where $\Theta = \{\theta = (\delta, \alpha, \beta, \gamma_1, \ldots, \gamma_p)^T : 0 < \delta_L \leq \delta \leq \delta_U, \epsilon \leq \alpha + \beta \leq 1 - \epsilon, v_i \leq \gamma_i \leq V_i, i = 1, \ldots, p\}$ for some $\epsilon > 0$ and $v_i, V_i \in \mathbb{R}$ for $i = 1, \ldots, p$.

The theorem below establishes the consistency of the maximum likelihood estimates of the parameters in model (3.3.1).

Theorem 3.3.1. *Consider the Poisson INGARCH with covariates model (3.3.1), and assume the conditions in Proposition 3.3.1. Furthermore, assume that assumption*

(AC) holds. Then the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent, i.e.,

$$\hat{\theta}_n \to \theta_0, \quad as\ n \to \infty.$$

Proof. The proof takes a similar approach as in showing Theorem 2.3.1. We first establish the identifiability of the model and then shows the consistency by Lemma 2.3.1. Denote

$$l_t(\theta) = Y_t(Z_t^T\gamma) + Y_t \log \mu_t(\theta) - \mu_t(\theta)e^{Z_t^T\gamma}, \tag{3.3.12}$$

then according to (3.3.4), the log likelihood function of the model is $l(\theta) = \sum_{t=1}^n l_t(\theta)$. It is easy to see that $\mathrm{E}l_t(\theta) < \infty$, hence $\mathrm{E}l_t^+(\theta) < \infty$. Denote $M_n(\theta) = 1/n \sum_{t=1}^n l_t(\theta)$ and $M(\theta) = \mathrm{E}l_1(\theta)$, where the expectations are taken under the true parameter value θ_0. It then follows from Proposition 3.3.1 and the extended mean ergodic theorem (see e.g., Billingsley (1995) pp. 284 and 495) that $M_n(\theta) \to M(\theta)$, almost surely as $n \to \infty$. To show the identifiability, one needs to show that $M(\theta) < M(\theta_0)$ for any $\theta \neq \theta_0$. First note that by Jensen's inequality on conditional expectations,

$$\begin{aligned}
M(\theta) - M(\theta_0) &= \mathrm{E}[l_1(\theta) - l_1(\theta_0)] \\
&= \mathrm{E}\big\{\mathrm{E}[\log \frac{p(Y_1|e^{Z_1^T\gamma}\mu_1(\theta))}{p(Y_1|e^{Z_1^T\gamma_0}\mu_1(\theta_0))}|\mathcal{F}_0]\big\} \\
&\leq \mathrm{E}\big\{\log \mathrm{E}[\frac{p(Y_1|e^{Z_1^T\gamma}\mu_1(\theta))}{p(Y_1|e^{Z_1^T\gamma_0}\mu_1(\theta_0))}|\mathcal{F}_0]\big\} \\
&= \mathrm{E}\log(1) = 0,
\end{aligned}$$

where $p(k|x)$ is the probability mass function of a Poisson distribution with intensity x evaluated at k. Hence $M(\theta) \leq M(\theta_0)$ for all $\theta \in \Theta$. The equality holds if and only if $e^{Z_1^T\gamma}\mu_1(\theta) = e^{Z_1^T\gamma_0}\mu_1(\theta_0)$, a.s. given $\mathcal{F}_0$, and this can happen only when $\gamma = \gamma_0$ and

$\mu_1(\theta) = \mu_1(\theta_0)$, a.s. According to the dynamics (3.3.1), this implies that

$$\begin{aligned}(\beta e^{-Z_0^T\gamma_0} - \beta_0 e^{-Z_0^T\gamma_0})Y_0 &= \delta_0 - \delta + \alpha_0\{\frac{\delta_0}{1-\alpha_0} + \beta_0 \sum_{k=0}^{\infty} \alpha_0^k Y_{-1-k} e^{-Z_{-1-k}^T\gamma_0}\} \\ &\quad -\alpha\{\frac{\delta}{1-\alpha} + \beta_0 \sum_{k=0}^{\infty} \alpha^k Y_{-1-k} e^{-Z_{-1-k}^T\gamma_0}\}.\end{aligned}$$

Since $\mathrm{Var}(Y_t|\mathcal{F}_{t-1}) > 0$, it follows that $\beta = \beta_0$. Similarly, one can show that $\alpha = \alpha_0$ and $\delta = \delta_0$. Hence $M(\theta) < M(\theta_0)$ for all $\theta \neq \theta_0$, which implies the identifiability of the model. In addition, one can show that $\mathrm{E}\sup_{\theta\in\Theta} l_t(\theta) < \infty$ under the conditions of the theorem. Hence according to Lemma 2.3.1 and following the similar arguments employed in showing Theorem 2.3.1, the result follows. □

The following theorem addresses the asymptotic distribution of the MLE and the idea of the proof is similar to that in Davis *et al.* (2003) and Theorem 2.3.2. For clear and easy notation, denote $W_t(\theta) = \log \mu_t(\theta) + Z_t^T\gamma$. Without otherwise indicated, W_t and $\dot{W}_t$ are both evaluated at θ_0, i.e., $W_t = W_t(\theta_0)$ and $\dot{W}_t = (\partial W_t(\theta)/\partial\theta)|_{\theta=\theta_0}$.

Theorem 3.3.2. *Consider the Poisson INGARCH with covariates (3.3.1) and assume the conditions in Proposition 3.3.1. Furthermore, assume that assumption (AC) holds. Then the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically normal, i.e.,*

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

where $\Omega = E\{e^{W_1}\dot{W}_1\dot{W}_1^T\}$.

Proof. According to (3.3.12), $l_t(\theta) = Y_t W_t(\theta) - e^{W_t(\theta)}$. We define a linearized form of $W_t(\theta)$ as $W^\dagger(\theta) = W_t + (\theta - \theta_0)^T\dot{W}_t$, and the corresponding linearized log-likelihood function of $l(\theta)$ as

$$l^\dagger(\theta) := \sum_{t=1}^{n} Y_t W_t^\dagger(\theta) - e^{W_t^\dagger(\theta)}.$$

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

$$\begin{aligned} R_n^\dagger(u) &:= l^\dagger(\theta_0) - l^\dagger(\theta_0 + un^{-1/2}) \\ &= -u^T n^{-1/2} \sum_{t=1}^{n} Y_t \dot{W}_t + \sum_{t=1}^{n} e^{W_t}(e^{u^t n^{-1/2}\dot{W}_t} - 1) \\ &= u^T n^{-1/2} \sum_{t=1}^{n} (Y_t - e^{W_t})\dot{W}_t \\ &\quad + \sum_{t=1}^{n} e^{W_t}(e^{u^T n^{-1/2}\dot{W}_t} - u^T n^{-1/2}\dot{W}_t - 1). \end{aligned} \tag{3.3.13}$$

The first term in (3.3.13) can be written as $u^T H_n$ where $H_n = n^{-1/2}\sum_{t=1}^{n}(Y_t - e^{W_t})\dot{W}_t$. Denote $s_t = n^{-1/2}(Y_t - e^{W_t})\dot{W}_t$, then it follows from the fact that $\mathrm{E}(s_t|\mathcal{F}_{t-1}) = n^{-1/2}\dot{W}_t \mathrm{E}\{Y_t - e^{W_t}|\mathcal{F}_{t-1}\} = 0$ that $\{s_t\}$ is a sequence of martingale difference sequences. Note that

$$\begin{aligned} \sum_{t=1}^{n} \mathrm{E}(s_t s_t^T|\mathcal{F}_{t-1}) &= \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}\{(Y_t - e^{W_t})^2 \dot{W}_t \dot{W}_t^T|\mathcal{F}_{t-1}\} \\ &= \frac{1}{n}\sum_{t=1}^{n} e^{W_t}\dot{W}_t\dot{W}_t^T, \end{aligned}$$

which converges almost surely to Ω by the mean ergodic theorem guaranteed by Proposition 3.3.1, provided that $\mathrm{E}|e^{W_t}\dot{W}_t\dot{W}_t^T| < \infty$. To see this, it suffices to show that $\|e^{W_t/2}\dot{W}_{t,i}\| < \infty$ for $i = 1, \ldots, p+3$, where $\|\cdot\|$ is the L^2-norm. We only give the derivation with respect to $\theta_2 = \alpha$ and $\theta_4 = \gamma_1$, since the proofs for δ, β and $\gamma_2, \ldots, \gamma_p$ are similar. First note that $\partial W_t(\theta)/\partial\alpha = 1/\mu_t(\theta)\partial\mu_t(\theta)/\partial\alpha$, then it follows from

(3.3.6) that

$$\begin{aligned}
\|e^{W_t/2}\dot{W}_{t,2}\|^2 &= \mathrm{E}\left|e^{Z_t^T\gamma_0}\mu_t(\theta_0)\frac{1}{\mu_t^2(\theta_0)}\left\{\frac{\partial\mu_t(\theta)}{\partial\alpha}|_{\theta=\theta_0}\right\}^2\right| \\
&= \mathrm{E}\left|e^{Z_t^T\gamma_0}\frac{1}{\mu_t(\theta_0)}\left\{\frac{\partial\mu_t(\theta)}{\partial\alpha}|_{\theta=\theta_0}\right\}^2\right| \\
&\leq C_1\mathrm{E}\left\{\frac{\partial\mu_t(\theta)}{\partial\alpha}|_{\theta=\theta_0}\right\}^2,
\end{aligned}$$

due to assumption (Z1) and the fact that $\mu_t(\theta_0) \geq \delta_0/(1-\alpha_0)$, where C_1 is a constant dependent on θ_0. According to (3.3.6), we have $\partial\mu_t(\theta)/\partial\alpha = \sum_{k=1}^{\infty}\alpha^{k-1}\mu_{t-k}(\theta)$. One can show that under the assumptions of the theorem that $\|\mu_t(\theta)\| < \infty$. Hence $\|\partial\mu_t(\theta)/\partial\alpha\| \leq \sum_{k=1}^{n}\alpha^{k-1}\|\mu_t(\theta)\| < \infty$, which implies that $\|e^{W_t/2}\dot{W}_{t,2}\| < \infty$. As for $\theta_4 = \gamma_1$, we have

$$\begin{aligned}
\|e^{W_t/2}\dot{W}_{t,4}^2\|^2 &= \mathrm{E}\left|e^{Z_t^T\gamma_0}\mu_t(\theta_0)\left\{\frac{1}{\mu_t(\theta_0)}\frac{\partial\mu_t(\theta)}{\partial\gamma_1}|_{\theta=\theta_0} + Z_{t,1}\right\}^2\right| \\
&= \mathrm{E}\left|e^{Z_t^T\gamma_0}\frac{1}{\mu_t(\theta_0)}\left\{\frac{\partial\mu_t(\theta_0)}{\partial\gamma_{1,0}}\right\}^2 + e^{Z_t^T\gamma_0}\mu_t(\theta_0)Z_{t,1}^2 + 2e^{Z_t^T\gamma_0}Z_{t,1}\frac{\partial\mu_t(\theta_0)}{\partial\gamma_{1,0}}\right|,
\end{aligned}$$

where $\partial\mu_t(\theta_0)/\partial\gamma_{1,0} = \partial\mu_t(\theta)/\partial\gamma_1|_{\theta=\theta_0}$. So it is sufficient to show that

$$\mathrm{E}\left|\frac{\partial\mu_t(\theta_0)}{\partial\gamma_{1,0}}\right| < \infty \quad \text{and} \quad \mathrm{E}\left\{\frac{\partial\mu_t(\theta_0)}{\partial\gamma_{1,0}}\right\}^2 < \infty.$$

Note that according to (3.3.6), we have $\partial\mu_t(\theta)/\partial\gamma_1 = -\beta\sum_{k=1}^{\infty}\alpha^{k-1}Y_{t-k}e^{-Z_{t-k}^T\gamma}Z_{t-k,1}$, so there exists a constant C_2 such that $|\partial\mu_t(\theta)/\partial\gamma_1| \leq C_2\sum_{k=1}^{\infty}\alpha^{k-1}Y_{t-k}$. Hence

$\mathrm{E}\left|\partial\mu_t(\theta_0)/\partial\gamma_{1,0}\right|^i < \infty$ for $i = 1, 2$. Moreover, for any $\epsilon > 0$,

$$\begin{aligned}
\sum_{t=1}^{n} \mathrm{E}\left\{s_t s_t^T \mathbf{1}_{[|s_t|\geq\epsilon]}|\mathcal{F}_{t-1}\right\} &= 1/n\sum_{t=1}^{n}\dot{W}_t\dot{W}_t^T\mathrm{E}\left\{(Y_t - e^{W_t})^2\mathbf{1}_{[|(Y_t-e^{W_t})\dot{W}_t|\geq\epsilon\sqrt{n}]}|\mathcal{F}_{t-1}\right\} \\
&\leq 1/n\sum_{t=1}^{n}\dot{W}_t\dot{W}_t^T\mathrm{E}\left\{(Y_t - e^{W_t})^2\mathbf{1}_{[|(Y_t-e^{W_t})\dot{W}_t|\geq M]}|\mathcal{F}_{t-1}\right\} \\
&\overset{n\to\infty}{\longrightarrow} \mathrm{E}\left\{(Y_t - e^{W_t})^2\dot{W}_t\dot{W}_t^T\mathbf{1}_{[|(Y_t-e^{W_t})\dot{W}_t|\geq M]}|\mathcal{F}_{t-1}\right\} \\
&\longrightarrow 0, \quad \text{as } M \to \infty.
\end{aligned}$$

Hence it follows from the central limit theorem for martingale difference sequences that

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

By Taylor expansion, the second term in (3.3.13) is

$$u^T\left\{\frac{1}{2n}\sum_{t=1}^{n} e^{W_t}\dot{W}_t\dot{W}_t^T\right\}u + \mathcal{O}_p(n^{-3/2}\sum_{t=1}^{n} e^{W_t}(u^T\dot{W}_t)^3),$$

in which the first term converges to $u^T\Omega u/2$ and the second term converges to zero. Hence $R_n^\dagger(u) \xrightarrow{\mathcal{L}} -u^TV + u^T\Omega u/2$, where $V \sim N(0, \Omega)$. It therefore follows that $\mathrm{argmin}_u R_n^\dagger(u) \xrightarrow{\mathcal{L}} \mathrm{argmin}_u\{-u^TV + u^T\Omega u/2\} = \Omega^{-1}V \sim N(0, \Omega^{-1})$.

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

$$\begin{aligned}
R_n^\dagger(u) - R_n(u) &= \sum_{t=1}^{n}(Y_t - e^{W_t})\left\{W_t(\theta) - W_t - (\theta-\theta_0)^T\dot{W}_t\right\} \\
&\quad - \sum_{t=1}^{n}\big[e^{W_t(\theta)} - e^{W_t + u^Tn^{-1/2}\dot{W}_t} \\
&\quad - e^{W_t}\{W_t(\theta) - W_t - (\theta-\theta_0)^T\dot{W}_t\}\big]. \qquad (3.3.14)
\end{aligned}$$

The first term of (3.3.14) is

$$\begin{aligned} A_n &= \sum_{t=1}^{n}(Y_t - e^{W_t})\{W_t(\theta) - W_t - n^{-1/2}u^T\dot{W}_t\} \\ &= u^T\left[\frac{1}{2n}\sum_{t=1}^{n}(Y_t - e^{W_t})\left\{\ddot{W}_t + (\ddot{W}_t(\theta^*) - \ddot{W}_t)\right\}\right]u, \end{aligned}$$

where $\ddot{W}_t = \partial^2 W_t(\theta)/\partial\theta^2|_{\theta=\theta_0}$ and θ^* lies between θ and θ_0. Since $\mathrm{E}\{(Y_t - e^{W_t})\ddot{W}_t\} = 0$, so it follows from the mean ergodic theorem that $A_n \to 0$ uniformly for $|u| \le K$, for all $K < \infty$ provided that $\ddot{W}_t(\theta^*) - \ddot{W}_t \xrightarrow{P} 0$. The second term in (3.3.14) is

$$B_n = -\sum_{t=1}^{n}\big[e^{W_t(\theta)} - e^{W_t + u^T n^{-1/2}\dot{W}_t} - e^{W_t}\{W_t(\theta) - W_t - (\theta-\theta_0)^T\dot{W}_t\}\big],$$

which, after Taylor expansion on the terms e^{W_t}, $e^{u^T n^{-1/2}\dot{W}_t}$ and $W_t(\theta)$, becomes

$$\begin{aligned} B_n &= -\sum_{t=1}^{n}\left[e^{W_t} + u^T n^{-1/2}e^{W_t}\dot{W}_t + u^T\frac{1}{2n}e^{W_t(\theta_1^*)}\{W_t^2(\theta_1^*) + \ddot{W}_t(\theta_1^*)\}u\right] \\ &\quad -\sum_{t=1}^{n}e^{W_t}\{1 + u^T n^{-1/2}\dot{W}_t + e^c\frac{1}{2n}u^T\dot{W}_t{}^2 u\} \\ &\quad -\sum_{t=1}^{n}e^{W_t}\{u^T\frac{1}{2n}\ddot{W}_t(\theta_2^*)u\}, \end{aligned}$$

where $0 \le c \le u^T\dot{W}_t/(2n)$ and $\|\theta_i^* - \theta_0\| \le \|\theta - \theta_0\|$ for $i = 1, 2$. Assuming each term in the above expression converges to a finite quantity in probability, we have that $B_n \to 0$ uniformly on compacts sets for u. Therefore $R_n^\dagger(u) - R_n(u) \to 0$ uniformly for $|u| \le K$, which implies $\mathrm{argmin}_u R_n(u)$ and $\mathrm{argmin}_u R_n^\dagger(u)$ have the same asymptotic distribution, i.e.,

$$\mathrm{argmin}_u R_n(u) \xrightarrow{\mathcal{L}} \Omega^{-1}V \sim N(0, \Omega^{-1}).$$

Note that $\mathrm{argmin}_u R_n(u) = \mathrm{argmax}_u l(\theta_0 + un^{-1/2}) = \sqrt{n}(\hat{\theta}_n - \theta_0)$, where $\hat{\theta}_n$ is the maximum likelihood estimator. Hence

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

□

3.3.3 Data Application

The Poisson INGARCH (3.3.1) and negative binomial INGARCH (3.3.2) models with covariates are fitted to three data examples. The estimates of the parameters are obtained by optimizing the corresponding log likelihood function using a Newton-Raphson method.

1. Number of road crashes in Schiphol

We consider the number of daily road crashes in Schiphol area in the Netherlands for the year 2001. It is of practical interest in accidents analysis to investigate the risk impact of traffic exposure and weather conditions. Real traffic exposure is typically defined as the total amount of vehicle kilometers driven on the major road network of a specific city region. However, in the absence of such data, earlier research has shown that weekday/weekend dummy variables can serve as a proxy (Levine *et al.* (1995a), Levine *et al.* (1995b), Brijs *et al.* (2008)). Weather conditions are measured in terms of daily temperatures in degrees Celsius. Figure 3.3 depicts the number of road crashes, autocorrelation of the observations and the temperature.

Models (3.3.1) and (3.3.2) are fitted to the data, where the covariates include the standardized temperature and a binary variable with values 0 and 1 representing weekday and weekend, respectively (see also Pedeli and Karlis (2010)). Table 3.4 summarizes the estimates of the parameters, whose standard deviations are calculated according to the Hessian matrices (3.3.7) and (3.3.11). Not surprisingly, the results indicate that the incidence of road crashes is larger during weekdays or when the temperature is lower.

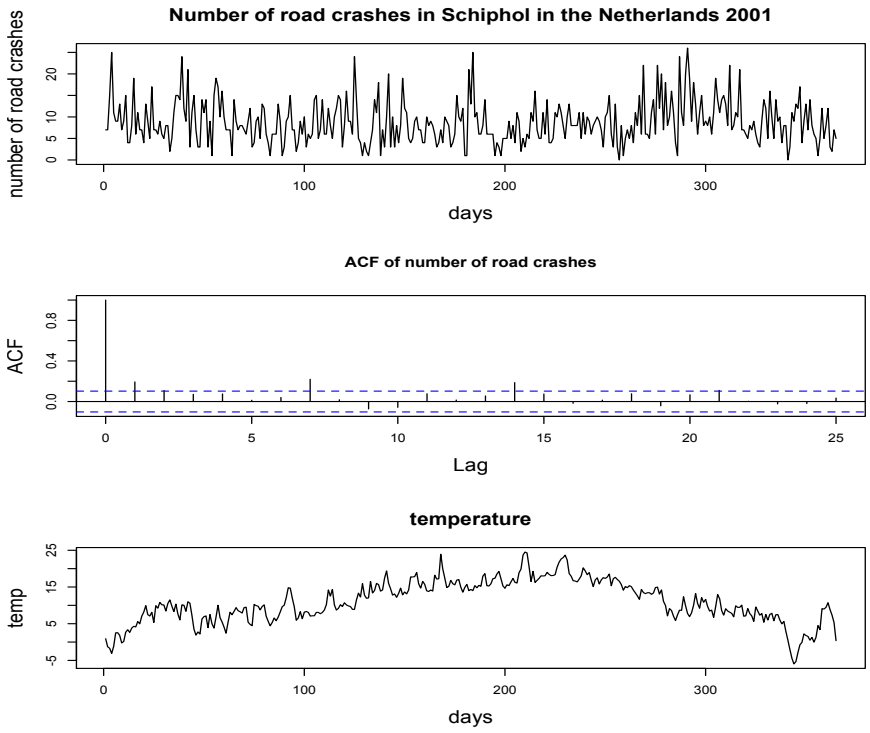

Figure 3.3: Top: Daily number of road crashes in Schiphol. Middle: ACF of the observations. Bottom: Temperature in degrees Celsius.

Table 3.4: Estimation of INGARCH with covariates on road crashes in Schiphol

	Poisson-based			NB-based with $r = 7$		
	estimate	s.e.	p-value	estimate	s.e.	p-value
δ	3.2495	0.9433	$< .01$	3.2069	1.3616	$< .01$
α	0.5086	0.1107	$< .01$	0.5027	0.1603	$< .01$
β	0.1581	0.0310	$< .01$	0.1690	0.0475	$< .01$
temperature	-0.0597	0.0239	0.013	-0.0568	0.0364	0.012
weekday/weekend	-0.4313	0.0419	$< .01$	-0.4413	0.0602	$< .01$

Table 3.5: Quantitative model checking for Schiphol road crashes data

Model	log likelihood	BIC	p-value of PIT	LS	QS	RPS
Poisson w/ covariates	-1124.44	2278.39	$< 10^{-5}$	3.0833	-0.0629	2.5477
NB w/ covariates	**−1039.73**	**2114.86**	0.7118	**2.8500**	**−0.0714**	**2.4695**
Poisson w/o covariates	-1182.87	2383.45	$< 10^{-5}$	3.2439	-0.0514	2.7724
NB w/o covariates	-1065.14	2153.90	0.8210	2.9195	-0.0643	2.6506

For comparison, Poisson INGARCH and negative binomial INGARCH models without covariates are also fitted to the data. Figure 3.4 plots the fitted conditional mean processes of the two negative binomial-based models. It appears from the graph that the model with covariates is more capable of capturing the fluctuation of the time series. To assess the goodness of fit and measure prediction power, an array of tools, both graphically and quantitatively, is utilized. Figure 3.5 and Figure 3.6 demonstrate the results of ACF of Pearson residuals and the randomized PIT test. Table 3.5 summarizes log likelihood functions, information criteria and various prediction scores (see (2.5.1), (2.5.2) and (2.5.3)). It can be seen that both negative binomial-based models, with or without covariates, pass the PIT test, while only the Pearson residuals of the negative binomial-based model with covariates appear to be white, and the prediction scores of it are consistently the smallest.

2. Number of polio incidences in the US

As another illustration, the polio data, which was first studied by Zeger (1988) and consists of monthly count of the number of poliomyelitis cases in US from 1970 to 1983 reported by the Centers for Disease Control, is investigated. It serves as a benchmark data set in the field of time series of counts. The observations with sample size $n = 168$, as depicted in Figure 3.7, range from 0 to 14, with a sample mean and variance of 1.33 and 3.5, respectively.

As indicated in Zeger (1988) and Jung and Tremayne (2011), significant serial

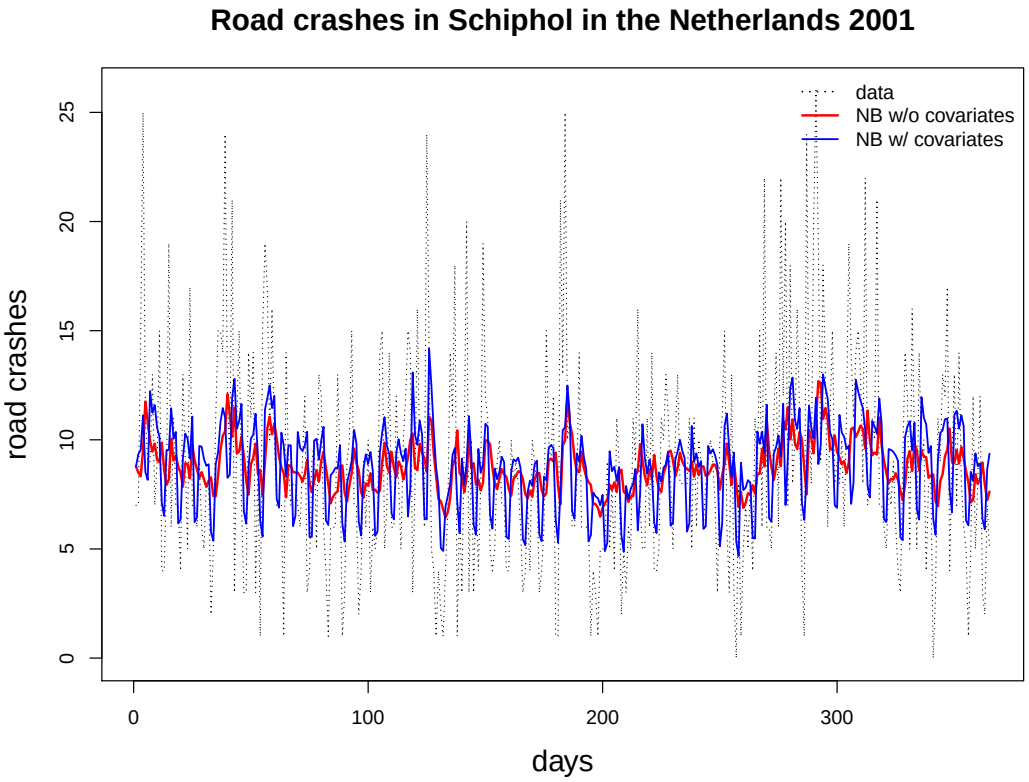

Figure 3.4: Fitted conditional mean processes of negative binomial-based INGARCH with and without covariates. The dotted curve is the observations, the dashed one is the fitted conditional mean process using the model without covariates, and the solid is the one using the model with covariates.

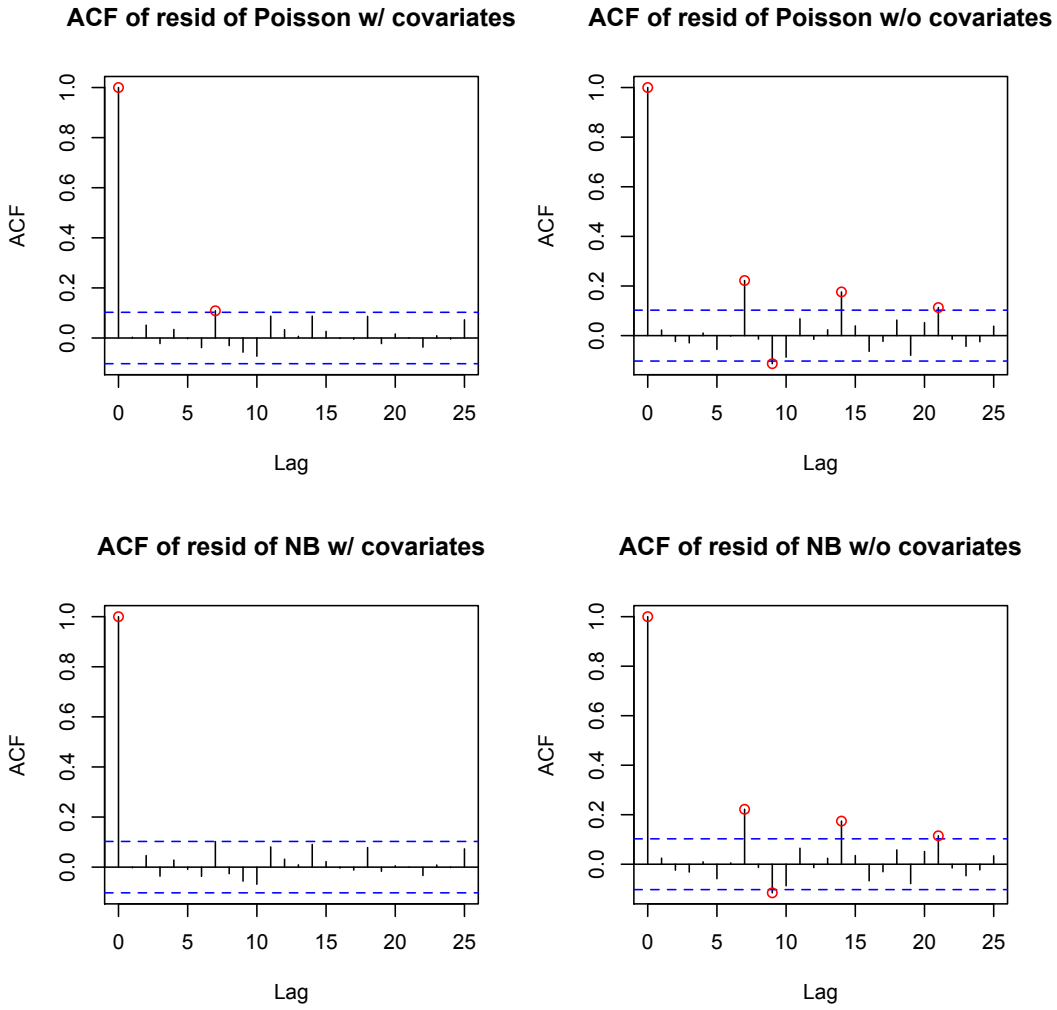

Figure 3.5: ACF of the Pearson residuals of all of the four models fitted to the Schiphol data. Circles in the plots correspond to the lags that have significant autocorrelations.

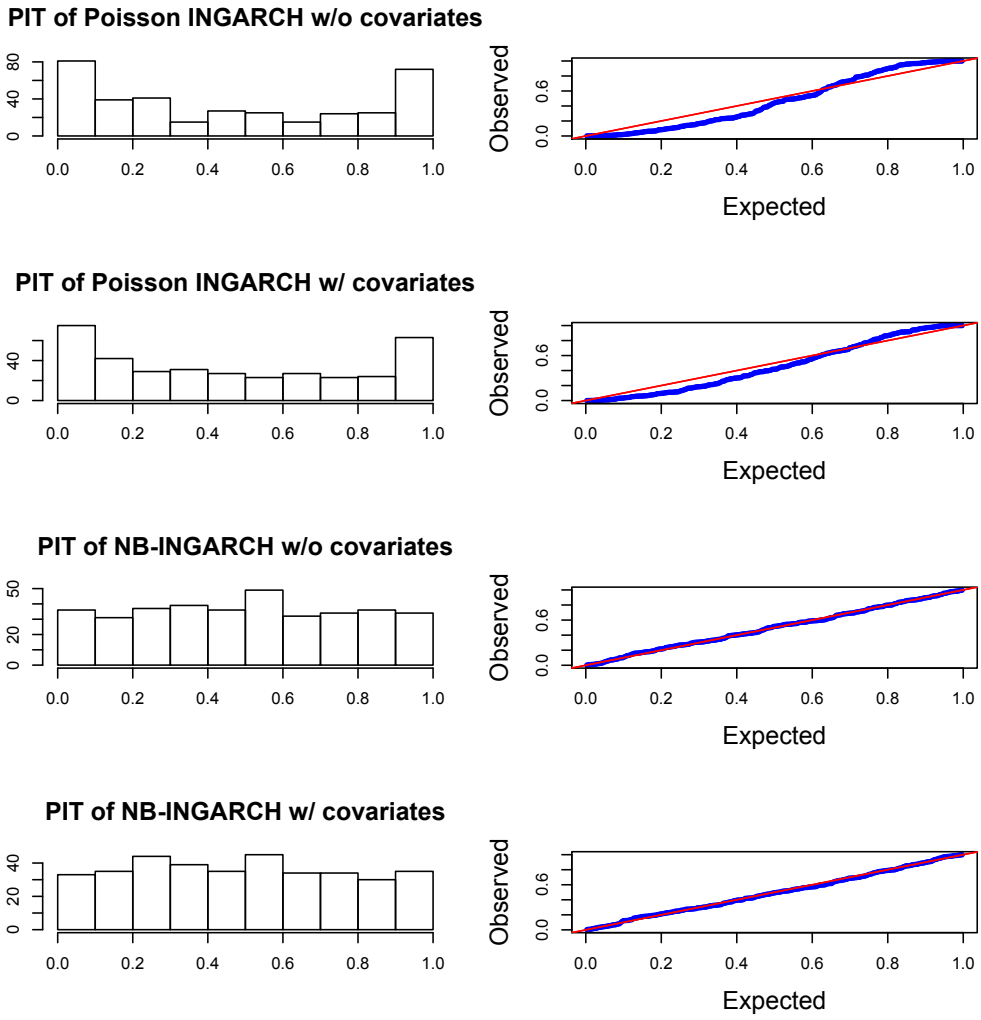

Figure 3.6: Randomized PIT test of four models fitted to the Schiphol data: Poisson and negative binomial-based INGARCH with and without covariates. Left: histograms of randomized PIT. Right: Q-Q plots of the PIT values versus Unif$(0, 1)$ distribution.

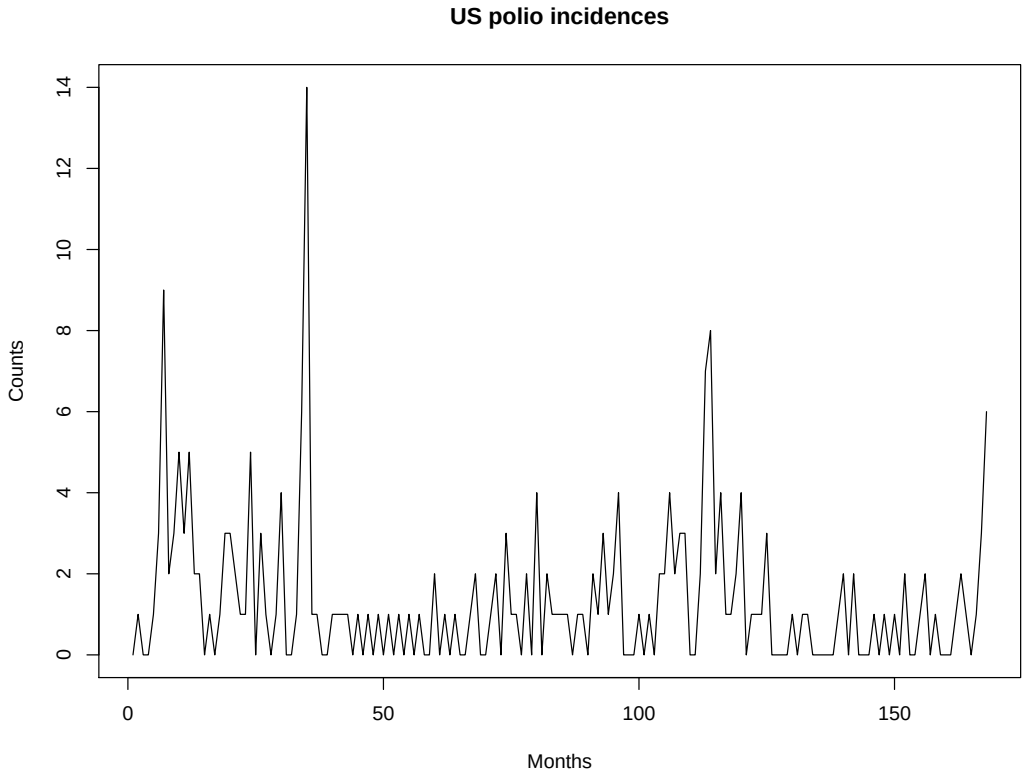

Figure 3.7: Number of polio incidences in the US, referred to as the *polio data*.

dependence and seasonal behavior can be observed in the data. Moreover, Zeger (1988) originally concerned with detecting a downward trend in the observations. To allow for the possible trend and seasonal components, models (3.3.1) and (3.3.2) with covariates

$$Z_t = (t/1000, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))^T$$

are considered. Both Poisson and negative binomial-based INGARCH models without covariates are also fitted for comparison. Table 3.6 summarizes the estimation results of the Poisson and negative binomial-based INGARCH with covariates, in which the standard errors and p-values are calculated based on the empirical Fisher information. It provides evidence to the existence of a downward trend and seasonal components. Figure 3.8 depicts the fitted conditional mean processes from two NB-INGARCH models. It appears that the NB-INGARCH with covariates fits the data set better than the model without covariates. Moreover, some diagnostic tools are implemented to compare the fitted models, as summarized in Table 3.7. It shows that the NB-INGARCH with covariates outperforms all the other models and has consistently the smallest prediction scores.

3. Number of asthma presentations in an Australian hospital

The last application considered in this section is the asthma data set. First studied by Davis *et al.* (2000), it consists of four-year daily counts of patients presenting at the accident and emergency department of a Campbelltown hospital located in the south-west metropolitan area of Sydney, Australia from year 1990 to 1993. The observations are plotted in Figure 3.9 by years. Davis *et al.* (2003) carried out a comprehensive treatment to choose explanatory regression variables, including air pollution, seasonal effects and the possible impact of the terms in the K-12 school year. The same set

Table 3.6: Estimation of INGARCH with covariates on polio data

	Poisson-based			NB-based with $r = 2$		
	estimate	s.e.	p-value	estimate	s.e.	p-value
δ	0.9861	0.381	$< .01$	0.8226	0.490	0.093
α	0.2160	0.204	0.291	0.3177	0.296	0.283
β	0.2531	0.066	$< .01$	0.2295	0.085	$< .01$
$t/1000$	-5.5266	2.098	$< .01$	-5.038	2.589	0.052
$\sin(2\pi t/6)$	0.0339	0.100	0.736	0.0528	0.133	0.691
$\cos(2\pi t/6)$	0.4679	0.104	$< .01$	0.4177	0.135	$< .01$
$\sin(2\pi t/12)$	-0.5100	0.142	$< .01$	-0.4532	0.176	$< .01$
$\cos(2\pi t/12)$	0.1997	0.114	0.08	0.1952	0.153	0.203

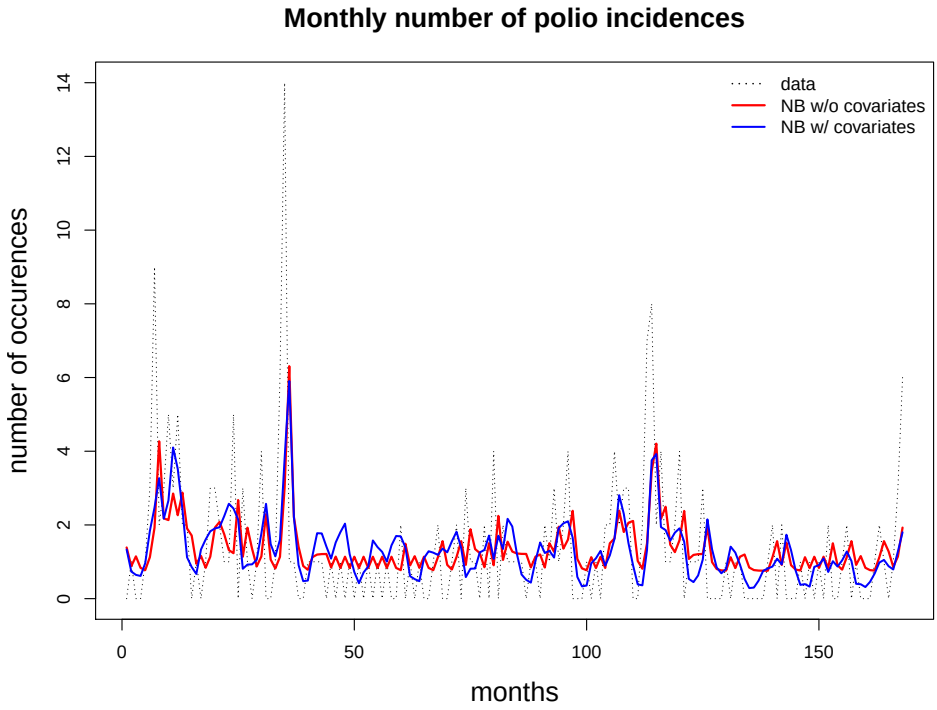

Figure 3.8: Dotted curve: true observations; Dashed curve: fitted conditional mean using NB-INGARCH; Solid curve: fitted conditional mean using NB-INGARCH with covariates.

Table 3.7: Quantitative model checking for polio data

Model	log likelihood	AIC	p-value of PIT	LS	QS	RPS
Poisson INGARCH	-279.37	564.75	0.0040	1.665	-0.253	0.830
NB INGARCH	-257.52	523.05	0.8664	1.536	-0.269	0.799
1-knot Poisson model	-279.28	570.56	0.0351	1.664	-0.256	0.831
1-knot NB model	-257.33	526.65	0.7987	1.535	-0.271	0.801
Poisson w/ covariates	-260.72	537.44	0.3076	1.553	-0.271	0.762
NB w/ covariates	$\mathbf{-247.81}$	**513.6**	0.8543	**1.478**	$\mathbf{-0.284}$	**0.739**

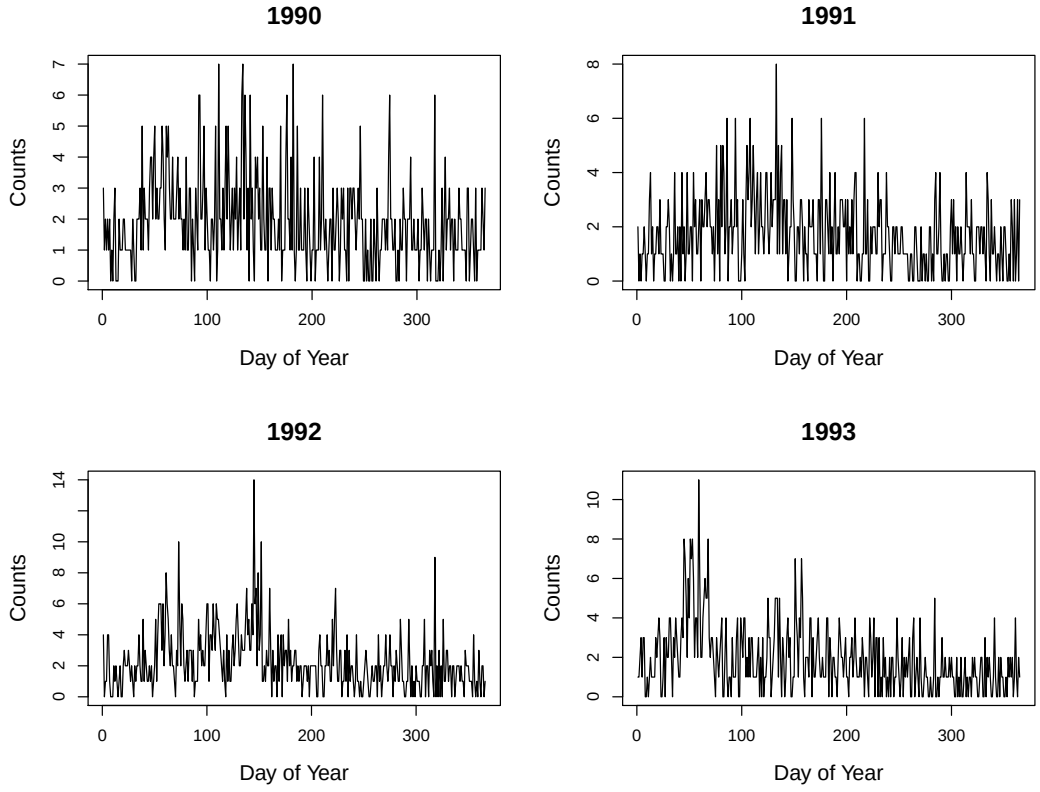

Figure 3.9: Number of asthma presentations in an Australian hospital from year 1990 to year 1993

Table 3.8: Estimation of INGARCH with covariates on asthma data

	Poisson-based			NB-based with $r = 25$		
	estimate	s.e.	p-value	estimate	s.e.	p-value
δ	0.2769	0.239	0.247	0.2559	0.256	0.318
α	0.8200	0.143	$< .01$	0.8348	0.152	$< .01$
β	0.0250	0.017	0.133	0.0218	0.017	0.192
Sunday	0.1986	0.052	$< .01$	0.1957	0.054	$< .01$
Monday	0.2273	0.051	$< .01$	0.2294	0.053	$< .01$
Annual Cosine	-0.2149	0.042	$< .01$	-0.2120	0.043	$< .01$
Annual Sine	0.1685	0.044	$< .01$	0.1707	0.046	$< .01$
Humidity	0.0087	0.003	$< .01$	0.0086	0.003	$< .01$
NO_2	-0.1037	0.034	$< .01$	-0.1027	0.035	$< .01$

of covariates is adopted in our research, which includes two annual harmonic terms $\cos(2\pi t/365)$ and $\sin(2\pi t/365)$, the Sunday and Monday effects, lagged composite humidity variable ($H_t/20$), NO_2 measurements and terms 1 and 2 for all four years. See Davis *et al.* (2000) and Davis *et al.* (2003) for details of the variable definitions and selection.

Four models: Poisson and negative binomial-based INGARCH with and without covariates are fitted to this data set. The MLE of the parameters are presented in Table 3.8. The signs of the corresponding covariates are consistent with the results given in Davis *et al.* (2003). Figure 3.10 depicts the fitted conditional mean processes using the two NB-based models for all the four years. It appears that the model with covariates is more capable of capturing the structure and fluctuations of the observations. A summary of the quantitative goodness of fit and model checking is provided in Table 3.9, which indicates that the models with covariates, whether Poisson or negative binomial based, are favored compared to the models without covariates.

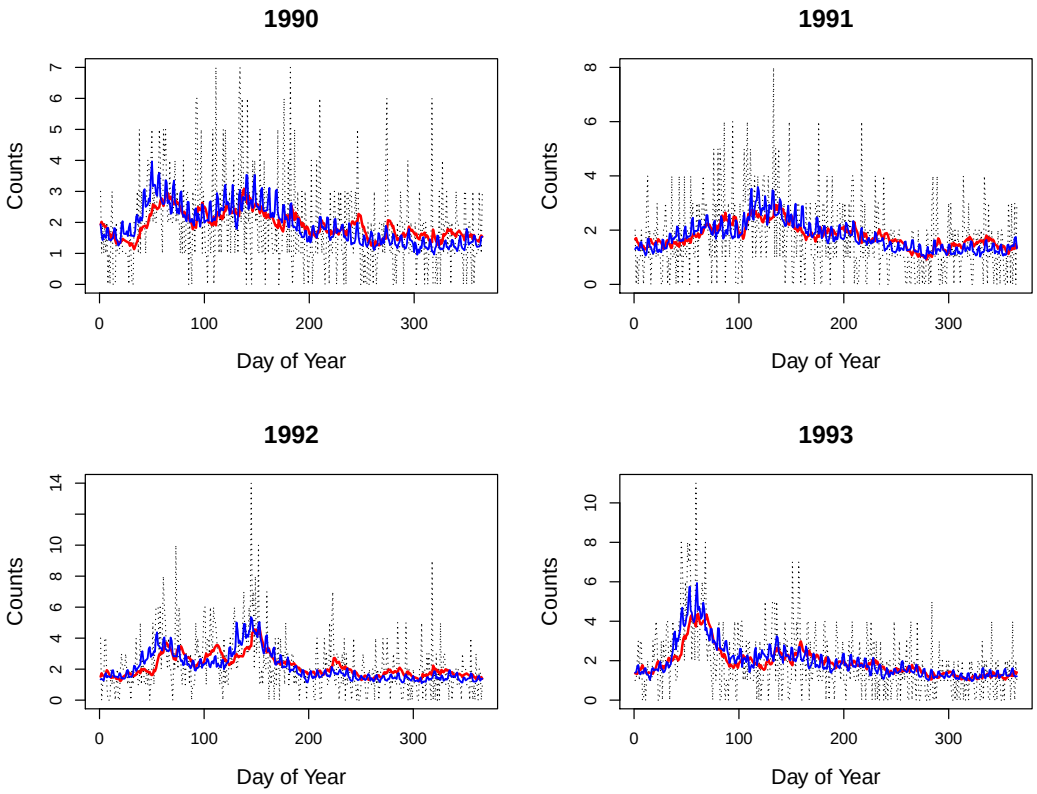

Figure 3.10: Dotted curve: true observations; Dashed curve: fitted conditional mean using NB-INGARCH; Solid curve: fitted conditional mean using NB-INGARCH with covariates.

Table 3.9: Quantitative model checking for asthma data

Model	log likelihood	AIC	p-value of PIT	LS	QS	RPS
Poisson INGARCH	-2490.6	4987.1	0.232	1.705	-0.215	0.795
NB INGARCH	-2481.8	4971.6	0.866	1.698	-0.215	0.794
Poisson w/ covariates	-2422.9	**4879.9**	0.808	**1.658**	**−0.222**	**0.757**
NB w/ covariates	**−2422.0**	4880.1	0.758	**1.658**	**−0.222**	**0.757**

3.4 Models of Orders Beyond One

3.4.1 Model Formulation

We extend the base model proposed in Chapter 2 to higher orders (p, q), in which the conditional mean X_t is allowed to depend on previous p conditional means and previous q observations. Specifically, the model is defined as

$$Y_t|\mathcal{F}_{t-1} \sim p(y|\eta_t), \quad X_t = g_\theta(X_{t-1}, \ldots, X_{t-p}, Y_{t-1}, \ldots, Y_{t-q}), \tag{3.4.1}$$

where $X_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$, $p, q \in \mathbb{N}$ and θ is the parameter vector. Here g_θ is a non-negative function defined on $[0, \infty)^p \times \mathbb{N}_0^q$ when Y_t are non-negative integers, or on $[0, \infty)^{p+q}$ when Y_t has a continuous conditional distribution. Throughout, we assume that the function g_θ satisfies a contraction condition, i.e., for any $x_1, \ldots, x_p, x'_1, \ldots, x'_p \geq 0$ and $y_1, \ldots, y_q, y'_1, \ldots, y'_q \in [0, \infty)$ or $\mathbb{N}_0$,

$$|g_\theta(x_1, \ldots, x_p, y_1, \ldots, y_q) - g_\theta(x'_1, \ldots, x'_p, y'_1, \ldots, y'_q)| \leq \sum_{i=1}^{p} a_i|x_i - x'_i| + \sum_{j=1}^{q} b_j|y_j - y'_j|, \tag{3.4.2}$$

where $a_1, \ldots, a_p, b_1, \ldots, b_q$ are non-negative constants with $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$. Note that model (3.4.1) includes the Poisson integer-valued GARCH(p, q) as a special case.

3.4.2 Stability Properties

Again we will take advantage of the IRF approach to investigate stability properties of the model, and the concepts of GMC and τ-weak dependence both play a critical role in establishing the relevant results. Here we suppress θ and use g to denote g_θ. Set $Z_t = (X_t, \ldots, X_{t-p+1})$ if $q = 1$, and $Z_t = (X_t, \ldots, X_{t-p+1}, Y_{t-1}, \ldots, Y_{t-q+1})$ if $q > 1$. It shows in the following proposition that $Z_t = f_{U_t}(Z_{t-1})$ for some well-defined random function f, where the sequence $\{U_t\}$ follows Uniform$(0, 1)$ independently. We point

out that the different definitions of Z_t for different values of q is necessary, in order to comply with the requirement that $\{U_t\}$ should be independent in the construction of an iterated random functions system.

Proposition 3.4.1. *Assume model (3.4.1) with the function g satisfying the contraction condition (3.4.2), and Z_t is defined as above. Then*

(a) $\{Z_t\}$ is a GMC Markov chain and has a unique stationary distribution π.

(b) Let $\{Z_t\}$ be the stationary process with $Z_1 \sim \pi$, then it is τ-weakly dependent, hence is an ergodic process.

Proof. To prove (a), we first consider the case that $q = 1$, and according to the model formulation, $X_t = g(X_{t-1}, \ldots, X_{t-p}, Y_{t-1})$. It is easy to see that $Z_t = (X_t, \ldots, X_{t-p+1})$ is a Markov chain. In order to apply the IRF approach, for $u \in (0, 1)$, the random function $f_u(z) : [0, \infty)^p \to [0, \infty)^p$ is defined as

$$f_u(z) := (g(x_1, \ldots, x_p, F^{-1}_{x_1}(u)), x_1, \ldots, x_{p-1}),$$

where $z = (x_1, \ldots, x_p) \in [0, \infty)^p$, F_x is the cumulative distribution function (CDF) of the one-parameter exponential family with mean x, and $F^{-1}_x(u) = \inf\{t \geq 0 : F_x(t) \geq u\}$. It suffices to verify Theorem 6.3.2. The norm on the state space E is defined as $\|z\| = \sum_{i=1}^p \omega_i |x_i|$, where $\omega_1, \ldots, \omega_p > 0$ are yet to be determined. For any $z_0 = (x^0_1, \ldots, x^0_p) \in E$, $\mathrm{E}\|z_0 - f_u(z_0)\| = \omega_1 \int_0^1 |x^0_1 - g(x^0_1, \ldots, x^0_p, F^{-1}_{x^0_1}(u))| du + \sum_{i=2}^p \omega_i |x^0_i - x^0_{i-1}| < \infty$. Next for $z_0 \in E$ fixed and any $z = (x_1, \ldots, x_p) \in E$, it remains to show that there exists $r \in (0, 1)$ such that $\mathrm{E}\|f_u(z) - f_u(z_0)\| \leq r\|z - z_0\|$,

where

$$\begin{aligned}
\mathrm{E}\|Z_1(z) - Z_1(z_0)\| &= \omega_1 \int_0^1 |g(x_1, \ldots, x_p, F_{x_1}^{-1}(u)) - g(x_1^0, \ldots, x_p^0, F_{x_1^0}^{-1}(u))| du \\
&\quad + \omega_2 |x_1 - x_1^0| + \ldots + \omega_p |x_{p-1} - x_{p-1}^0| \\
&\leq [\omega_1(a_1 + b) + \omega_2] |x_1 - x_1^0| + (\omega_1 a_2 + \omega_3) |x_2 - x_2^0| \\
&\quad + \ldots + (\omega_1 a_{p-1} + \omega_p) |x_{p-1} - x_{p-1}^0| + \omega_1 a_p |x_p - x_p^0|, \quad (3.4.3)
\end{aligned}$$

which follows from the stochastic monotonicity of the one-parameter exponential family. By comparing the corresponding coefficients on both sides of (3.4.3), the inequality $\mathrm{E}\|Z_1(z) - Z_1(z_0)\| \leq r\|z - z_0\|$ proves to be equivalent to $\omega_1(a_1 + b) + \omega_2 \leq r\omega_1, \omega_1 a_2 + \omega_3 \leq r\omega_2, \ldots, \omega_1 a_{p-1} + \omega_p \leq r\omega_{p-1}$ and $\omega_1 a_p \leq r\omega_p$ hold at the same time for some $r \in (0, 1)$. Denote $h_1(r) := r^p - (a_1 + b)r^{p-1} - a_2 r^{p-2} - \ldots - a_{p-1} r - a_p$, then the above inequalities can be reduced to $h_1(r_1) \geq 0$ for some $r_1 \in (0, 1)$. If $a_1 + \ldots + a_p + b < 1$, then $h_1(0) = -a_p < 0$ and $h_1(1) = 1 - a_1 - \ldots - a_p - b > 0$, so there must exist $r_1 \in (0, 1)$ such that $h_1(r_1) \geq 0$. The values of $\omega_1, \ldots, \omega_p$ can be determined accordingly. It follows from induction that $\mathrm{E}\|Z_n(z) - Z_n(z_0)\| \leq r^n \|z - z_0\|$ for all $n \geq 1$. Hence $\{Z_t, t \geq 1\}$ is geometric moment contracting, and has a unique stationary distribution.

Now we consider the case that $q \geq 2$ and the Markov chain is defined as $Z_t = (X_t, \ldots, X_{t-p+1}, Y_{t-1}, \ldots, Y_{t-q+1})$. For any $u \in (0, 1)$, define the random function $f_u(z) : [0, \infty)^p \times \mathbb{N}_0^{q-1} \to [0, \infty)^p \times \mathbb{N}_0^{q-1}$ (when Y_t is discrete) or $[0, \infty)^{p+q-1} \to [0, \infty)^{p+q-1}$ (when Y_t is continuous) to be

$$f_u(z) = (g(x_1, \ldots, x_p, F_{x_1}^{-1}(u), y_2, \ldots, y_q), x_1, \ldots, x_{p-1}, F_{x_1}^{-1}(u), y_2, \ldots, y_{q-1}),$$

where $z = (x_1, \ldots, x_p, y_2, \ldots, y_q)$. The norm is defined in the same way as above. Similarly to the case that $q = 1$, the inequalities implied by the two conditions in Theorem 6.3.2 are reduced to $h_2(r_2) \geq 0$ for some $r_2 \in (0, 1)$, where $h_2(r)$ is a polyno-

mial of order $\max\{p, q\}$. It turns out such r_2 exists provided $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$. Hence $\{Z_t\}$ is geometric moment contracting and has a unique stationary distribution.

According to the conditions in Theorem 6.4.1 and the uniqueness of the stationary distribution of $\{Z_t\}$, it follows that $\{Z_t\}$ is τ-weakly dependent, hence is a stationary and ergodic process. □

Chapter 4

Bivariate Poisson Autoregression

4.1 Introduction

In many applications, for example, in epidemiology, biology and accidents analysis, one often encounters multivariate count data. To this end, we consider the problem of modeling multivariate time series of counts. The main hurdle in this work is to choose an appropriate discrete distribution. Unlike in the continuous case, there is not one natural choice to model the contemporaneous dependence among multiple time series of counts. In the Gaussian case, one merely needs to substitute the multivariate normal distribution to achieve this goal. There has been a few attempts to model bivariate time series of counts, for example, Heinen and Rengifo (2003), Pedeli and Karlis (2010) and Pedeli and Karlis (2011). However, most of them are based upon thinning ideas or parameter-driven models. In this chapter, our aim is to extend the class of observation-driven models developed in Chapter 2 to multivariate count data.

In this chapter, we focus on the Poisson case and formulate a bivariate Poisson integer-valued GARCH (BINGARCH) model. This model is capable of modeling

the serial dependence between two time series of counts. Considering the difficulty exhibited in deriving the stability properties in the univariate INGARCH case (e.g., Propositions 2.2.3 and 2.2.4), it is expected that the establishment of these properties for the bivariate process is even more involved. Fortunately, the iterated random functions approach allows us to derive the stability properties under a contracting constraint on the coefficient matrices. In addition, the generalization of the BINGARCH model to higher orders is considered and the relevant theory is developed. Inference procedures are also presented and applied to a real data application in the area of traffic accident analysis.

The organization of this chapter is as follows. Section 4.2 defines a bivariate Poisson distribution and proposes the BINGARCH model. The stability properties of the model is also demonstrated in this section. Section 4.3 extends the model to higher orders. Inference based on the likelihood of these models is discussed in Section 4.4. The chapter concludes with an application to the number of daytime and nighttime road accidents near the Schiphol airport in the Netherlands, and comparisons between the univariate and bivariate model fitting are presented in Section 4.5.

4.2 Model Formulation and Stability Theory

For ease of discussion, only a bivariate model is investigated in this paper. However, the generalization to the multivariate case is possible using similarly stylized arguments. Denote $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2})^T$ as the bivariate observations at time t, that is, $\{Y_{t,1}, t \geq 1\}$ and $\{Y_{t,2}, t \geq 1\}$ are the two time series under consideration. A Poisson-based bivariate INGARCH (BINGARCH) model of order $(1,1)$ is defined as

$$\mathbf{Y}_t|\mathcal{F}_{t-1} \sim \mathrm{BP}(\lambda_{t,1}, \lambda_{t,2}, \phi),\ \boldsymbol{\lambda}_t = (\lambda_{t,1}, \lambda_{t,2})^T = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1}, \qquad (4.2.1)$$

where $\mathcal{F}_t = \sigma\{\boldsymbol{\lambda}_1, \mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$, $\phi \geq 0$, $\boldsymbol{\delta} = (\delta_1, \delta_2) \in \mathbb{R}_+^2$ and $\mathbf{A}, \mathbf{B}$ are both 2×2 matrices with nonnegative entries. The notation $\mathbf{Y}_t | \mathcal{F}_{t-1} \sim \text{BP}(\lambda_{t,1}, \lambda_{t,2}, \phi)$ represents the bivariate Poisson distribution whose probability mass function (pmf) is given by

$$\begin{aligned} P(Y_{t,1} = m, Y_{t,2} = n | \mathcal{F}_{t-1}) &= e^{-(\lambda_{t,1}+\lambda_{t,2}-\phi)} \frac{(\lambda_{t,1}-\phi)^m}{m!} \frac{(\lambda_{t,2}-\phi)^n}{n!} \\ &\quad \times \sum_{s=0}^{\min\{m,n\}} \binom{m}{s} \binom{n}{s} s! \Big(\frac{\phi}{(\lambda_{t,1}-\phi)(\lambda_{t,2}-\phi)} \Big)^s, \quad (4.2.2) \end{aligned}$$

where $\phi \in [0, \min\{\lambda_{t,1}, \lambda_{t,2}\})$. The definition (4.2.2) allows for modeling dependence between $Y_{t,1}$ and $Y_{t,2}$, and $\text{Cov}(Y_{t,1}, Y_{t,2} | \mathcal{F}_{t-1}) = \phi$. In fact, there exist independent random variables $X_1 \sim \text{Pois}(\lambda_{t,1} - \phi)$, $X_2 \sim \text{Pois}(\lambda_{t,2} - \phi)$ and $X_3 \sim \text{Pois}(\phi)$ such that $Y_{t,1} = X_1 + X_3$ and $Y_{t,2} = X_2 + X_3$. For a comprehensive treatment of a multivariate Poisson distribution, readers can refer to Kocherlakota and Kocherlakota (1992) and Johnson *et al.* (1997). Model (4.2.1) is capable of capturing dependence between the two time series $\{Y_{t,1}\}$ and $\{Y_{t,2}\}$, provided that the parameter $\phi \neq 0$, or the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ are not both diagonal. Note that by recursion, for any $l \geq 1$, we have $\boldsymbol{\lambda}_t = (\mathbf{I} + \mathbf{A} + \ldots + \mathbf{A}^{l-1})\boldsymbol{\delta} + \mathbf{A}^l \boldsymbol{\lambda}_{t-l} + \sum_{k=0}^{l-1} \mathbf{A}^k \mathbf{B} \mathbf{Y}_{t-k-1}$, where $\mathbf{I}$ is the identity matrix. If $\rho(\mathbf{A}) < 1$, i.e., the largest absolute eigenvalue of $\mathbf{A}$ is less than 1, then

$$\begin{aligned} \boldsymbol{\lambda}_t &= (\mathbf{I} + \mathbf{A} + \mathbf{A}^2 \ldots)\boldsymbol{\delta} + \sum_{k=0}^{\infty} \mathbf{A}^k \mathbf{B} \mathbf{Y}_{t-k-1} \\ &= (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\delta} + \sum_{k=0}^{\infty} \mathbf{A}^k \mathbf{B} \mathbf{Y}_{t-k-1}. \quad (4.2.3) \end{aligned}$$

Hence under the condition that $\rho(\mathbf{A}) < 1$, we have $\boldsymbol{\lambda}_t \geq (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\delta}$ for all t. This provides a feasible upper bound for parameter ϕ in practice, since $\phi \leq \min\{\lambda_{t,1}, \lambda_{t,2}\}$ for all t according to (4.2.2).

Before stating the main result, we introduce some relevant notation for a general matrix $\mathbf{J} \in \mathbb{C}^{m \times n}$. Define $\|\mathbf{J}\|_p$ as the p-induced norm of matrix $\mathbf{J}$ for $1 \leq p \leq \infty$, i.e., $\|\mathbf{J}\|_p = \max_{\mathbf{x} \neq \mathbf{0}} \{\|\mathbf{J}\mathbf{x}\|_p / \|\mathbf{x}\|_p : \mathbf{x} \in \mathbb{C}^n\}$, where $\|\mathbf{x}\|_p$ is the p-norm of the vector $\mathbf{x}$. In

particular, $\|\mathbf{J}\|_1$ is the maximum absolute column sum of $\mathbf{J}$, $\|\mathbf{J}\|_\infty$ is the maximum absolute row sum, and $\|\mathbf{J}\|_2$ is the square root of its largest singular value if $\mathbf{J}$ is a square matrix. Note that $\rho(\mathbf{J}) \leq \|\mathbf{J}\|_p$ for any $1 \leq p \leq \infty$, where the spectral radius $\rho(\mathbf{J})$ is the largest absolute eigenvalue of $\mathbf{J}$. If $\mathbf{J}$ is diagonal, then $\rho(\mathbf{J}) = \|\mathbf{J}\|_1 = \|\mathbf{J}\|_\infty$.

The study focuses on the bivariate Markov chain $\{\boldsymbol{\lambda}_t, t \geq 1\}$. According to the pmf of a bivariate Poisson distribution (4.2.2), the random function $f_{\mathbf{u}}(\boldsymbol{\lambda})$ for $\mathbf{u} = (u_1, u_2, u_3) \in [0,1]^3$ is defined as

$$f_{\mathbf{u}}(\boldsymbol{\lambda}) = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda} + \mathbf{B}\widetilde{F}^{-1}_{\mathbf{u}}(\boldsymbol{\lambda}), \tag{4.2.4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$, $\widetilde{F}^{-1}_{\mathbf{u}}(\boldsymbol{\lambda}) = (F^{-1}_{\lambda_1-\phi}(u_1) + F^{-1}_{\phi}(u_3), F^{-1}_{\lambda_2-\phi}(u_2) + F^{-1}_{\phi}(u_3))^T \in \mathbb{N}_0^2$, and $F^{-1}_x(u) = \inf\{t \geq 0 : F_x(t) \geq u\}$. Hence it can be seen that for all t, $\boldsymbol{\lambda}_t = f_{\mathbf{U}_t}(\boldsymbol{\lambda}_{t-1})$, where $\{\mathbf{U}_t, t \geq 1\}$ follow independent uniform distribution on $[0,1]^3$.

Proposition 4.2.1. *Assume model (4.2.1), and $\boldsymbol{\delta}, \mathbf{A}$ and $\mathbf{B}$ have nonnegative entries.*

(a) If $\rho(\mathbf{A}+\mathbf{B}) < 1$, then there exists at least one stationary distribution to $\{\boldsymbol{\lambda}_t\}$. In addition, if $\|\mathbf{A}\|_p < 1$ for some $1 \leq p \leq \infty$, then the stationary distribution is unique.

(b) If $\|\mathbf{A}\|_p + 2^{(1-1/p)}\|\mathbf{B}\|_p < 1$ for some $1 \leq p \leq \infty$, then $\{\boldsymbol{\lambda}_t\}$ is a GMC Markov chain with a unique stationary and ergodic distribution, denoted by π.

Proposition 4.2.1 (a) provides a weaker condition than (b) to guarantee the existence of a unique stationary distribution, but does not yield ergodicity. Note that if all of the entries of $\mathbf{A}$ and $\mathbf{B}$ are nonnegative, then $\rho(\mathbf{A}) \leq \rho(\mathbf{A}+\mathbf{B})$. To see this, note that for any $k \geq 1$, $\|\mathbf{A}^k\|_1^{1/k} \leq \|(\mathbf{A}+\mathbf{B})^k\|_1^{1/k}$. Then by virtue of Gelfand's formula (Gelfand (1941)), we have $\rho(\mathbf{A}) = \lim_{k\to\infty} \|\mathbf{A}^k\|_1^{1/k} \leq \lim_{k\to\infty} \|(\mathbf{A}+\mathbf{B})^k\|_1^{1/k} = \rho(\mathbf{A}+\mathbf{B})$.

Proof. First note that $\{\lambda_t\}$ is a weak Feller chain, i.e., $Pf \in C_b(E)$ for any $f \in C_b(E)$, where $C_b(E)$ is the set of bounded continuous functions defined on the state space $E = [0, \infty) \times [0, \infty)$. To see this, for $\mathbf{x} = (x_1, x_2)$, we have $P_{\mathbf{x}} f = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} f(\boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_1 + \mathbf{B}(m\ n)^T | \boldsymbol{\lambda}_1 = \mathbf{x}) p(m, n|\mathbf{x})$, where $p(m, n|\mathbf{x})$ is the pmf of $\mathrm{BP}(x_1, x_2, \phi)$, so it follows from the continuity of f that $P_{\mathbf{x}} f$ is also continuous. Then according to Theorem 6.2.1, it suffices to show that $\{\boldsymbol{\lambda}_t\}$ is bounded in probability on average, i.e., for any $\boldsymbol{\lambda}_1 \in E$ and $\epsilon > 0$, there exists $C = [0, K_1] \times [0, K_2] \in \mathbb{R}^2$, such that $1/k \sum_{t=1}^{k} P^t(\boldsymbol{\lambda}_1, C) \geq 1 - \epsilon$ for all $k \geq 1$, where $P^t(\boldsymbol{\lambda}_1, \cdot)$ is the t-th transition probability of $\{\boldsymbol{\lambda}_t\}$. It follows from $\boldsymbol{\lambda}_2 = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_1 + \mathbf{B}\mathbf{Y}_1$ that $\mathrm{E}(\boldsymbol{\lambda}_2|\boldsymbol{\lambda}_1) = \boldsymbol{\delta} + (\mathbf{A} + \mathbf{B})\boldsymbol{\lambda}_1$. Then by induction, we have for any $t \geq 1$,

$$\mathrm{E}(\boldsymbol{\lambda}_{t+1}|\boldsymbol{\lambda}_1) = [\mathbf{I} + (\mathbf{A} + \mathbf{B}) + \ldots + (\mathbf{A} + \mathbf{B})^{t-1}]\boldsymbol{\delta} + (\mathbf{A} + \mathbf{B})^t \boldsymbol{\lambda}_1, \tag{4.2.5}$$

where $\mathbf{I}$ is a 2×2 identity matrix. It follows from $\rho(\mathbf{A} + \mathbf{B}) < 1$ that $(\mathbf{A} + \mathbf{B})^t \to \mathbf{0}$, as $t \to \infty$, $\mathbf{I} - (\mathbf{A} + \mathbf{B})$ is nonsingular and $[\mathbf{I} - (\mathbf{A} + \mathbf{B})]^{-1} = \sum_{t=0}^{\infty} (\mathbf{A} + \mathbf{B})^t$. So for all $t \geq 1$,

$$\begin{aligned} \mathrm{E}(\boldsymbol{\lambda}_{t+1}|\boldsymbol{\lambda}_1) &\leq (\mathbf{I} + (\mathbf{A} + \mathbf{B}) + (\mathbf{A} + \mathbf{B})^2 + \ldots)\boldsymbol{\delta} + (\mathbf{A} + \mathbf{B})\boldsymbol{\lambda}_1 \\ &= [\mathbf{I} - (\mathbf{A} + \mathbf{B})]^{-1}\boldsymbol{\delta} + (\mathbf{A} + \mathbf{B})\boldsymbol{\lambda}_1. \end{aligned}$$

It then follows that

$$\begin{aligned} P^t(\boldsymbol{\lambda}_{t+1} \in C|\boldsymbol{\lambda}_1) &= P(\lambda_{t+1,1} \leq K_1, \lambda_{t+1,2} \leq K_2|\boldsymbol{\lambda}_1) \\ &\geq 1 - P(\lambda_{t+1,1} > K_1|\boldsymbol{\lambda}_1) - P(\lambda_{t+1,2} > K_2|\boldsymbol{\lambda}_1) \\ &\geq 1 - \mathrm{E}(\lambda_{t+1,1}|\boldsymbol{\lambda}_1)/K_1 - \mathrm{E}(\lambda_{t+1,2}|\boldsymbol{\lambda}_1)/K_2 \\ &= 1 - \boldsymbol{\nu}^T \mathrm{E}(\boldsymbol{\lambda}_{t+1}|\boldsymbol{\lambda}_1) \\ &\geq 1 - \boldsymbol{\nu}^T \{[\mathbf{I} - (\mathbf{A} + \mathbf{B})]^{-1}\boldsymbol{\delta} + (\mathbf{A} + \mathbf{B})\boldsymbol{\lambda}_1\}, \end{aligned}$$

where $\boldsymbol{\nu} = (1/K_1 \quad 1/K_2)^T$. It is easy to see that there exist K_1 and $K_2 \in \mathbb{R}$ large enough such that $\boldsymbol{\nu}^T\{[\mathbf{I} - (\mathbf{A} + \mathbf{B})]^{-1} + (\mathbf{A} + \mathbf{B})\boldsymbol{\lambda}_1\} \leq \epsilon$, which in turn gives $P^t(\boldsymbol{\lambda}_{t+1} \in C|\boldsymbol{\lambda}_1) \geq 1-\epsilon$ for all $t \geq 1$. Hence $1/k \sum_{t=1}^{k} P^t(\boldsymbol{\lambda}_1, C) \geq 1-\epsilon$ for all $k \geq 1$, which proves the boundedness in probability on average of $\{\boldsymbol{\lambda}_t\}$. Therefore $\{\boldsymbol{\lambda}_t\}$ has at least one stationary distribution.

Now further assume that $\|\mathbf{A}\|_p < 1$ for some $p \in [1, \infty]$. Since for all $l \geq 1$, we have $\boldsymbol{\lambda}_{l+1} = (\mathbf{I} + \mathbf{A} + \ldots + \mathbf{A}^{l-1})\boldsymbol{\delta} + \mathbf{A}^l\boldsymbol{\lambda}_1 + \sum_{k=0}^{l-1} \mathbf{A}^k\mathbf{B}\mathbf{Y}_{l-k}$, so it is then clear that $(\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\delta}$ is a reachable state if $\mathbf{Y}_l = \mathbf{Y}_{l-1} = \ldots \mathbf{Y}_1 = \mathbf{0}$ for some $l \in \mathbb{N}$ large enough. By virtue of Theorem 6.2.2, one yet needs to show that $\{\boldsymbol{\lambda}_t\}$ is an e-chain, i.e., for any continuous function f with compact support defined on $[0, \infty) \times [0, \infty)$ and $\epsilon > 0$, there exists an $\eta > 0$ such that $|P^k_{\mathbf{x}_1} f - P^k_{\mathbf{z}_1} f| < \epsilon$, for $\|\mathbf{x}_1 - \mathbf{z}_1\| < \eta$ and all $k \geq 1$, where $\mathbf{x}_1 = (x_{1,1}, x_{1,2})^T$, $\mathbf{z}_1 = (z_{1,1}, z_{1,2})^T$, and $\|\cdot\|$ is some norm defined on $\mathbb{R}^2$. Without loss of generality, assume $|f| \leq 1$. Take ϵ' and η sufficiently small such that $\epsilon' + 8\eta/(1 - \|\mathbf{A}\|_p) < \epsilon$ and $|f(\mathbf{x}_1) - f(\mathbf{z}_1)| < \epsilon'$ whenever $\|\mathbf{x}_1 - \mathbf{z}_1\|_p < \eta$, for some $p \in [1, \infty]$. Then for the case $k = 1$,

$$
\begin{aligned}
|P_{\mathbf{x}_1} f - P_{\mathbf{z}_1} f| &= |\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\big[f(\boldsymbol{\delta} + \mathbf{A}\mathbf{x}_1 + \mathbf{B}\binom{m}{n})p(m,n|\mathbf{x}_1) - f(\boldsymbol{\delta} + \mathbf{A}\mathbf{z}_1 + \mathbf{B}\binom{m}{n})p(m,n|\mathbf{z}_1)\big]| \\
&\leq \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p(m,n|\mathbf{x}_1)|f(\boldsymbol{\delta} + \mathbf{A}\mathbf{x}_1 + \mathbf{B}\binom{m}{n}) - f(\boldsymbol{\delta} + \mathbf{A}\mathbf{z}_1 + \mathbf{B}\binom{m}{n})| \\
&\quad + \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} |p(m,n|\mathbf{x}_1) - p(m,n|\mathbf{z}_1)||f(\boldsymbol{\delta} + \mathbf{A}\mathbf{z}_1 + \mathbf{B}\binom{m}{n})| \\
&= I + II,
\end{aligned}
$$

where $p(m, n|\mathbf{x}_1)$ is the pmf of $\mathrm{BP}(x_{1,1}, x_{1,2}, \phi)$ given by (4.2.2) and $\phi \leq \min\{x_{1,1}, x_{1,2}\}$ is the covariance. Denoting $p(i|x)$ as the pmf of a univariate Poisson distribution with

intensity x evaluated at i, it follows from (3.2.3) that

$$\begin{aligned}
& \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}|p(m,n|\mathbf{x}_1)-p(m,n|\mathbf{z}_1)| \\
\leq\ & \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{s=0}^{m\wedge n}|p(m-s|x_{1,1}-\phi)p(n-s|x_{1,2}-\phi)-p(m-s|z_{1,1}-\phi)p(z_{1,2}-\phi)|p(s|\phi) \\
\leq\ & \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{s=0}^{m\wedge n}|p(m-s|x_{1,1}-\phi)-p(m-s|z_{1,1}-\phi)|p(n-s|x_{1,2}-\phi)p(s|\phi) \\
& +\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{s=0}^{m\wedge n}p(m-s|z_{1,1}-\phi)|p(n-s|x_{1,2}-\phi)-p(n-s|z_{1,2}-\phi)|p(s|\phi) \\
=\ & \sum_{s=0}^{\infty}\sum_{n=s}^{\infty}\sum_{m=s}^{\infty}|p(m-s|x_{1,1}-\phi)-p(m-s|z_{1,1}-\phi)|p(n-s|x_{1,2}-\phi)p(s|\phi) \\
& +\sum_{s=0}^{\infty}\sum_{n=s}^{\infty}\sum_{m=s}^{\infty}p(m-s|z_{1,1}-\phi)|p(n-s|x_{1,2}-\phi)-p(n-s|z_{1,2}-\phi)|p(s|\phi) \\
\leq\ & \sum_{i=0}^{\infty}|p(i|x_{1,1}-\phi)-p(i|z_{1,1}-\phi)|+\sum_{i=0}^{\infty}|p(i|x_{1,2}-\phi)-p(i|z_{1,2}-\phi)| \\
\leq\ & 2(1-e^{-|x_{1,1}-z_{1,1}|})+2(1-e^{-|x_{1,2}-z_{1,2}|}).
\end{aligned}$$

Since $|x_{1,i}-z_{1,i}|\leq\|\mathbf{x}_1-\mathbf{z}_1\|_1\leq c_p\|\mathbf{x}_1-\mathbf{z}_1\|_p$, for $i=1,2$ and any $1\leq p\leq\infty$, where $c_p=2^{1-1/p}\leq 2$, so for any $\mathbf{x}_1,\mathbf{z}_1$ and $p\in[1,\infty]$, we have

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}|p(m,n|\mathbf{x}_1)-p(m,n|\mathbf{z}_1)|\leq 4(1-e^{-2\|\mathbf{x}_1-\mathbf{z}_1\|_p}). \tag{4.2.6}$$

So it follows from $|f|\leq 1$ that $II\leq 4(1-e^{-2\|\mathbf{x}_1-\mathbf{z}_1\|_p})$. Since $\|\boldsymbol{\delta}+\mathbf{A}\mathbf{x}_1+\mathbf{B}(m\ \ n)^T-(\boldsymbol{\delta}+\mathbf{A}\mathbf{z}_1+\mathbf{B}(m\ \ n)^T)\|_p=\|\mathbf{A}(\mathbf{x}_1-\mathbf{z}_1)\|_p\leq\|\mathbf{A}\|_p\|\mathbf{x}_1-\mathbf{z}_1\|_p\leq\eta$, so $I\leq\epsilon'$. Hence

$$|P_{\mathbf{x}_1}f-P_{\mathbf{z}_1}f|\leq\epsilon'+4(1-e^{-2\|\mathbf{x}_1-\mathbf{z}_1\|_p}). \tag{4.2.7}$$

For the case $k = 2$, we have

$$\begin{aligned}|P^2_{\mathbf{x}_1} f - P^2_{\mathbf{z}_1} f| &= |\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}[p(m,n|\mathbf{x}_1)P_{\mathbf{x}_2} f - p(m,n|\mathbf{z}_1)P_{\mathbf{z}_2} f]| \\ &\leq \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p(m,n|\mathbf{x}_1)|P_{\mathbf{x}_2} f - P_{\mathbf{z}_2} f| + \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}|p(m,n|\mathbf{x}_1) - p(m,n|\mathbf{z}_1)||P_{\mathbf{z}_2} f|,\end{aligned}$$

where $\mathbf{x}_2 = \boldsymbol{\delta} + \mathbf{A}\mathbf{x}_1 + \mathbf{B}(m \;\; n)^T$ and $\mathbf{z}_2 = \boldsymbol{\delta} + \mathbf{A}\mathbf{z}_1 + \mathbf{B}(m \;\; n)^T$. Since $\|\mathbf{x}_2 - \mathbf{z}_2\|_p = \|\mathbf{A}(\mathbf{x}_1 - \mathbf{z}_1)\|_p \leq \|\mathbf{A}\|_p\|\mathbf{x}_1 - \mathbf{z}_1\|_p \leq \eta$, so it follows from (4.2.7) that

$$\begin{aligned}|P^2_{\mathbf{x}_1} f - P^2_{\mathbf{z}_1} f| &\leq \epsilon' + 4(1 - e^{-2\|\mathbf{x}_2 - \mathbf{z}_2\|_p}) + 4(1 - e^{-2\|\mathbf{x}_1 - \mathbf{z}_1\|_p}) \\ &\leq \epsilon' + 4(1 - e^{-2\|\mathbf{A}\|_p\|\mathbf{x}_1 - \mathbf{z}_1\|_p}) + 4(1 - e^{-2\|\mathbf{x}_1 - \mathbf{z}_1\|_p}).\end{aligned}$$

Hence by induction, we have for any $k \geq 1$ that

$$\begin{aligned}|P^k_{\mathbf{x}_1} f - P^k_{\mathbf{z}_1} f| &\leq \epsilon' + 4\sum_{s=0}^{k-1}(1 - e^{-2\|\mathbf{A}\|_p^s\|\mathbf{x}_1 - \mathbf{z}_1\|_p}) \\ &\leq \epsilon' + 8\sum_{s=0}^{\infty}\|\mathbf{A}\|_p^s\|\mathbf{x}_1 - \mathbf{z}_1\|_p \\ &\leq \epsilon' + \frac{8\eta}{1 - \|\mathbf{A}\|_p} < \epsilon,\end{aligned}$$

which proves that $\{\boldsymbol{\lambda}_t\}$ is an e-chain. Therefore there exists a unique stationary distribution to $\{\boldsymbol{\lambda}_t\}$.

To prove (b), it suffices to verify Theorem 6.3.2. The first condition holds trivially. For the second one, consider $\boldsymbol{\lambda}_0 = (\lambda_1^0, \lambda_2^0)^T$ fixed and any $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ belonging to the state space, and use $\|\cdot\|_p$ as the norm on it. Then

$$\begin{aligned}\mathrm{E}\|\boldsymbol{\lambda}_1(\boldsymbol{\lambda}) - \boldsymbol{\lambda}_1(\boldsymbol{\lambda}_0)\|_p &= \int \|(\boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda} + \mathbf{B}\widetilde{F}^{-1}_{\boldsymbol{\lambda}}(\mathbf{u})) - (\boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_0 + \mathbf{B}\widetilde{F}^{-1}_{\boldsymbol{\lambda}_0}(\mathbf{u}))\|_p \, d\mathbf{u} \\ &\leq \|\mathbf{A}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\|_p + \int \|\mathbf{B}[\widetilde{F}^{-1}_{\boldsymbol{\lambda}}(\mathbf{u}) - \widetilde{F}^{-1}_{\boldsymbol{\lambda}_0}(\mathbf{u})]\|_p \, d\mathbf{u} \\ &\leq \|\mathbf{A}\|_p\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_p + \|\mathbf{B}\|_p \int \|\widetilde{F}^{-1}_{\boldsymbol{\lambda}}(\mathbf{u}) - \widetilde{F}^{-1}_{\boldsymbol{\lambda}_0}(\mathbf{u})\|_1 \, d\mathbf{u},\end{aligned}$$

which follows from the definition of induced norm of matrices and the fact that $\|\widetilde{F}_{\boldsymbol{\lambda}}^{-1}(\mathbf{u}) - \widetilde{F}_{\boldsymbol{\lambda}_0}^{-1}(\mathbf{u})\|_p \leq \|\widetilde{F}_{\boldsymbol{\lambda}}^{-1}(\mathbf{u}) - \widetilde{F}_{\boldsymbol{\lambda}_0}^{-1}(\mathbf{u})\|_1$ for any $p \geq 1$. Since $\widetilde{F}_{\boldsymbol{\lambda}}^{-1}(\mathbf{u}) - \widetilde{F}_{\boldsymbol{\lambda}_0}^{-1}(\mathbf{u}) = (F_{\lambda_1-\phi}^{-1}(u_1) + F_{\phi}^{-1}(u_3), F_{\lambda_2-\phi}^{-1}(u_2) + F_{\phi}^{-1}(u_3))^T - (F_{\lambda_1^0-\phi}^{-1}(u_1) + F_{\phi}^{-1}(u_3), F_{\lambda_2^0-\phi}^{-1}(u_2) + F_{\phi}^{-1}(u_3))^T = (F_{\lambda_1-\phi}^{-1}(u_1) - F_{\lambda_1^0-\phi}^{-1}(u_1), F_{\lambda_2-\phi}^{-1}(u_2) - F_{\lambda_2^0-\phi}^{-1}(u_2))^T$, so we have

$$\begin{aligned}\int \|\widetilde{F}_{\boldsymbol{\lambda}}^{-1}(\mathbf{u}) - \widetilde{F}_{\boldsymbol{\lambda}_0}^{-1}(\mathbf{u})\|_1 \, d\mathbf{u} &= \int |F_{\lambda_1-\phi}^{-1}(u_1) - F_{\lambda_1^0-\phi}^{-1}(u_1)| + |F_{\lambda_2-\phi}^{-1}(u_2) - F_{\lambda_2^0-\phi}^{-1}(u_2)| \, d\mathbf{u} \\ &= |\lambda_1 - \lambda_1^0| + |\lambda_2 - \lambda_2^0| = \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_1.\end{aligned}$$

Hence it follows from $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_1 \leq 2^{(1-1/p)}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_p$ that

$$\begin{aligned}\mathrm{E}\|f_{\mathbf{u}}(\boldsymbol{\lambda}) - f_{\mathbf{u}}(\boldsymbol{\lambda}_0)\|_p &\leq \|\mathbf{A}\|_p\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_p + \|\mathbf{B}\|_p\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_1 \\ &\leq \|\mathbf{A}\|_p\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_p + 2^{(1-1/p)}\|\mathbf{B}\|_p\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_p \\ &\leq (\|\mathbf{A}\|_p + 2^{(1-1/p)}\|\mathbf{B}\|_p)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_p.\end{aligned}$$

So if $\|\mathbf{A}\|_p + 2^{(1-1/p)}\|\mathbf{B}\|_p < 1$, then $\{\boldsymbol{\lambda}_t\}$ is geometric moment contracting, and hence has a unique stationary distribution. According to Theorem 6.4.1, it is also τ-weakly dependent and is an ergodic casual Bernoulli shift process, which completes the proof of the proposition. □

4.3 Extension to a BINGARCH(m, n) Model

We now generalize model (4.2.1) to a BINGARCH(m, n), where $m, n \in \mathbb{N}$. In particular, it is defined as

$$\mathbf{Y}_t|\mathcal{F}_{t-1} \sim \mathrm{BP}(\lambda_{t,1}, \lambda_{t,2}, \phi), \ \boldsymbol{\lambda}_t = \boldsymbol{\delta} + \sum_{i=1}^{m} \mathbf{A}_i\boldsymbol{\lambda}_{t-i} + \sum_{j=1}^{n} \mathbf{B}_j\mathbf{Y}_{t-j}, \tag{4.3.1}$$

where $\boldsymbol{\lambda}_t = (\lambda_{t,1}, \lambda_{t,2})^T$, $\mathcal{F}_0 = \sigma\{\boldsymbol{\lambda}_0, \ldots, \boldsymbol{\lambda}_{1-s}\}$, $\mathcal{F}_t = \sigma\{\boldsymbol{\lambda}_0, \ldots, \boldsymbol{\lambda}_{1-s}, \mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$, where $s = \max\{m, n\}$ and $\boldsymbol{\delta} > 0$, $\mathbf{A}_i, \mathbf{B}_j, i = 1, \ldots, m; j = 1, \ldots, n$ are 2×2 matrices

with nonnegative entries. It follows from a similar calculation in (4.2.3) that $\boldsymbol{\lambda}_t \geq (\mathbf{I} - \sum_{i=1}^m \mathbf{A}_i)^{-1}\boldsymbol{\delta}$ for any t, provided that $\rho(\sum_{i=1}^m \mathbf{A}_i) < 1$.

In order to apply the IRF approach, define

$$Z_t = \begin{cases} (\boldsymbol{\lambda}_t, \ldots, \boldsymbol{\lambda}_{t-m+1}), & \text{if } n = 1, \\ (\boldsymbol{\lambda}_t, \ldots, \boldsymbol{\lambda}_{t-m+1}, \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_{t-n+1}), & \text{if } n > 1. \end{cases} \tag{4.3.2}$$

It is easy to see that Z_t defined above is a Markov chain according to the dynamics in model (4.3.1).

Proposition 4.3.1. *Consider model (4.3.1) and assume that* $\sum_{i=1}^m \|\mathbf{A}_i\|_p + 2^{(1-1/p)} \sum_{j=1}^n \|\mathbf{B}_j\|_p < 1$, *for some* $p \in \{1, \ldots, \infty\}$. *Then* $\{Z_t\}$ *defined in (4.3.2) is geometric moment contracting and* τ*-weakly dependent. Hence* $\{(\mathbf{Y}_t, \boldsymbol{\lambda}_t)\}$ *is stationary and ergodic.*

Proof. The idea of the proof is similar to that of Proposition 3.4.1. When $n = 1$, for any $z = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$, where $\boldsymbol{\lambda}_i = (\lambda_{i,1}, \lambda_{i,2})^T$, and $\mathbf{u} = (u_1, u_2, u_3) \in (0,1)^3$, the iterated random function can be defined as

$$f_u(z) = (\boldsymbol{\delta} + \sum_{i=1}^m \mathbf{A}_i\boldsymbol{\lambda}_i + \mathbf{B}_1\widetilde{F}^{-1}_{\boldsymbol{\lambda}_1}(\mathbf{u}), \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{m-1}),$$

where $\widetilde{F}^{-1}_{\boldsymbol{\lambda}_1}(\mathbf{u}) = (F^{-1}_{\lambda_{1,1}-\phi}(u_1) + F^{-1}_{\phi}(u_3), F^{-1}_{\lambda_{1,2}-\phi}(u_2) + F^{-1}_{\phi}(u_3))^T$. In order to verify Theorem 6.3.2, define the norm on the state space as $\|z\| = \sum_{i=1}^m w_i\|\boldsymbol{\lambda}_i\|_p$ for some $p \in \mathbb{Z} \cup \{+\infty\}$, where $w_i > 0, i = 1, \ldots, m$ are yet to be specified. It is trivial to verify Condition 1. For Condition 2, fix $z_0 = (\boldsymbol{\lambda}^0_1, \ldots, \boldsymbol{\lambda}^0_m)$, then for any $z = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$

and $\mathbf{u} \in (0,1)^3$, we have

$$\begin{aligned}
\mathrm{E}\|Z_1(z) - Z_1(z_0)\| &\leq w_1 \|\sum_{i=1}^{m} \mathbf{A}_i(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_i^0)\|_p + w_1 \int \|\mathbf{B}_1(\widetilde{F}_{\boldsymbol{\lambda}_1}^{-1}(\mathbf{u}) - \widetilde{F}_{\boldsymbol{\lambda}_1^0}^{-1}(\mathbf{u}))\|_p d\mathbf{u} \\
&\quad + w_2\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^0\|_p + \ldots + w_m\|\boldsymbol{\lambda}_{m-1} - \boldsymbol{\lambda}_{m-1}^0\|_p \\
&\leq w_1 \sum_{i=1}^{m} \|\mathbf{A}_i\|_p \|\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_i^0\|_p + 2^{(1-1/p)} w_1 \|\mathbf{B}_1\|_p \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^0\|_p \\
&\quad + w_2\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^0\|_p + \ldots + w_m\|\boldsymbol{\lambda}_{m-1} - \boldsymbol{\lambda}_{m-1}^0\|_p.
\end{aligned}$$

The second condition of geometric moment contraction requires that there exists an $r \in (0,1)$ such that $\mathrm{E}\|Z_1(z) - Z_1(z_0)\| \leq r\|z - z_0\|_p$, which by comparing coefficients on both sides, can be shown to be equivalent to that the polynomial $h_1(r) = r^m - (\|\mathbf{A}_1\|_p + 2^{(1-1/p)}\|\mathbf{B}_1\|_p)r^{m-1} - \|\mathbf{A}_2\|_p r^{m-2} - \ldots - \|\mathbf{A}_{m-1}\|_p r - \|\mathbf{A}_m\|_p$ has a root in $(0,1)$. Since $h_1(0) < 0$ and $h_1(1) = 1 - \sum_{i=1}^{m} \|\mathbf{A}_i\|_p - 2^{(1-1/p)}\|\mathbf{B}_1\|_p$, so the existence of such a root is guaranteed if $h_1(1) > 0$, which gives that $\sum_{i=1}^{m} \|\mathbf{A}_i\|_p + 2^{(1-1/p)}\|\mathbf{B}_1\|_p < 1$. Hence according to Theorem 6.3.2, the rest of the proposition follows.

If $n > 1$, for $z = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ and $\mathbf{u} \in (0,1)^3$, the iterated random function can be defined as

$$f_{\mathbf{u}}(z) = (\boldsymbol{\delta} + \sum_{i=1}^{m} \mathbf{A}_i\boldsymbol{\lambda}_i + \mathbf{B}_1\widetilde{F}_{\boldsymbol{\lambda}_1}^{-1}(\mathbf{u}) + \sum_{j=2}^{n} \mathbf{B}_j\mathbf{y}_j, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{m-1}, \widetilde{F}_{\boldsymbol{\lambda}_1}^{-1}(\mathbf{u}), \mathbf{y}_2, \ldots, \mathbf{y}_{n-1}).$$

With a similar definition of the norm on the state space, it can be shown that the second condition of geometric moment contracting holds if the polynomial $h_2(r) = r^s - (\|\mathbf{A}_1\|_p + 2^{(1-1/p)}\|\mathbf{B}_1\|_p)r^{s-1} - \ldots - (\|\mathbf{A}_s\|_p + 2^{(1-1/p)}\|\mathbf{B}_s\|_p)$ has a root in $(0,1)$, where $s = \max\{m,n\}$, $\mathbf{A}_i = \mathbf{0}$ for $i = m+1, \ldots, s$ and $\mathbf{B}_i = \mathbf{0}$ for $i = n+1, \ldots, s$. Since $h_2(0) < 0$ and $h_2(1) = 1 - \sum_{i=1}^{m} \|\mathbf{A}_i\|_p - 2^{(1-1/p)} \sum_{j=1}^{n} \|\mathbf{B}_j\|_p$, so such a root exists if $\sum_{i=1}^{m} \|\mathbf{A}_i\|_p + 2^{(1-1/p)} \sum_{j=1}^{n} \|\mathbf{B}_j\|_p < 1$, which completes the proof. □

4.4 Likelihood Inference

We begin with the maximum likelihood estimates of the parameters. Let $Y_1, \ldots, Y_n$ be observations from model (4.2.1), in which, without loss of generality, $\mathbf{A}$ is assumed to be diagonal, that is, $\mathbf{A} = \text{diag}(\alpha_1, \alpha_2)$. Denote $\mathbf{B} = (\beta_{ij})_{i,j=1,2}$, then the parameter vector $\theta = (\delta_1, \delta_2, \alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \phi)^T$. The likelihood function conditional on $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \lambda_{1,2})^T$ and based on the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is therefore given by

$$\begin{aligned} L(\theta|\mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\lambda}_1) &= f(\mathbf{Y}_1|\theta, \boldsymbol{\lambda}_1) \prod_{t=2}^{n} f(\mathbf{Y}_t|\mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1}, \boldsymbol{\lambda}_1, \theta) \\ &= \prod_{t=1}^{n} e^{-(\lambda_{t,1}+\lambda_{t,2}-\phi)} \frac{(\lambda_{t,1}-\phi)^{Y_{t,1}}}{Y_{t,1}!} \frac{(\lambda_{t,2}-\phi)^{Y_{t,2}}}{Y_{t,2}!} \\ &\quad \times \sum_{s=0}^{Y_{t,1} \wedge Y_{t,2}} \binom{Y_{t,1}}{s} \binom{Y_{t,2}}{s} s! \left(\frac{\phi}{(\lambda_{t,1}-\phi)(\lambda_{t,2}-\phi)}\right)^s, \end{aligned}$$

and the log likelihood function, up to a constant free of θ, is

$$\begin{aligned} l(\theta) &= -\sum_{t=1}^{n} (\lambda_{t,1} + \lambda_{t,2} - \phi) + \sum_{t=1}^{n} Y_{t,1} \log(\lambda_{t,1} - \phi) + \sum_{t=1}^{n} Y_{t,2} \log(\lambda_{t,2} - \phi) \\ &\quad + \sum_{t=1}^{n} \log\Big\{ \sum_{s=0}^{Y_{t,1} \wedge Y_{t,2}} \binom{Y_{t,1}}{s} \binom{Y_{t,2}}{s} s! \Big(\frac{\phi}{(\lambda_{t,1}-\phi)(\lambda_{t,2}-\phi)}\Big)^s \Big\}. \end{aligned} \tag{4.4.1}$$

Hence the maximum likelihood estimator is a solution to the constrained optimization problem, in which $l(\theta)$ is maximized subject to the constraint that $\phi \leq \min\{\delta_1/(1-\alpha_1), \delta_2/(1-\alpha_2)\}$. This constraint ensures that $\phi \leq \min\{\lambda_{t,1}, \lambda_{t,2}\}$ for all t according to the remark after (4.2.3). In addition, the estimates should also satisfy the conditions specified in Proposition 4.2.1 to guarantee the stability properties of the fitted model.

4.5 Data Application

A data set containing the daytime and nighttime number of road accidents in Schiphol area in the Netherlands is considered here (see also Pedeli and Karlis (2011)). The daytime accidents occurred between 10:00am and 6:00pm, while the nighttime accidents are classified as those occurred during the rest of the day. In accidents analysis, nighttime accidents happen usually due to different reasons from daytime accidents. For example, people are more likely to travel for entertainment or consume alcohol at night. These could be contributing factors for the seen difference between the daytime and nighttime accidents. Nevertheless, the numbers of daytime and nighttime accidents in the same region may have serial dependence, since they share similar environmental conditions, including weather conditions and characteristics of the road. The data are shown in Figure 4.1 with a correlation of 0.145 between the two time series. The autocorrelation and cross-correlation of them are provided in Figure 4.2. In particular, note that the time series of daytime accidents exhibits a strong seasonal effect with period 7.

A BINGARCH(1, 1) model was fitted to the daytime and nighttime accidents. However, based on the residual analysis, the model is inadequate since it is incapable of capturing the strong seasonal component. To overcome this limitation, a BINGARCH(7, 7) is considered, in which the dynamics are given by

$$\boldsymbol{\lambda}_t = \boldsymbol{\delta} + \mathbf{A}_1\boldsymbol{\lambda}_{t-1} + \mathbf{A}_7\boldsymbol{\lambda}_{t-7} + \mathbf{B}_1\mathbf{Y}_{t-1} + \mathbf{B}_7\mathbf{Y}_{t-7}.$$

In order to reduce model complexity, both $\mathbf{A}_1$ and $\mathbf{A}_7$ are taken to be diagonal matrices. The fitted model provides evidence of significant correlation between the two time series with $\hat{\phi} = 0.33$, $\hat{\mathbf{A}}_1 = \text{diag}\{0, 0.49\}$ and $\hat{\mathbf{A}}_7 = \text{diag}\{0.49, 0.23\}$. Figure 4.3 plots the fitted conditional mean processes and ACF of the Pearson residuals of

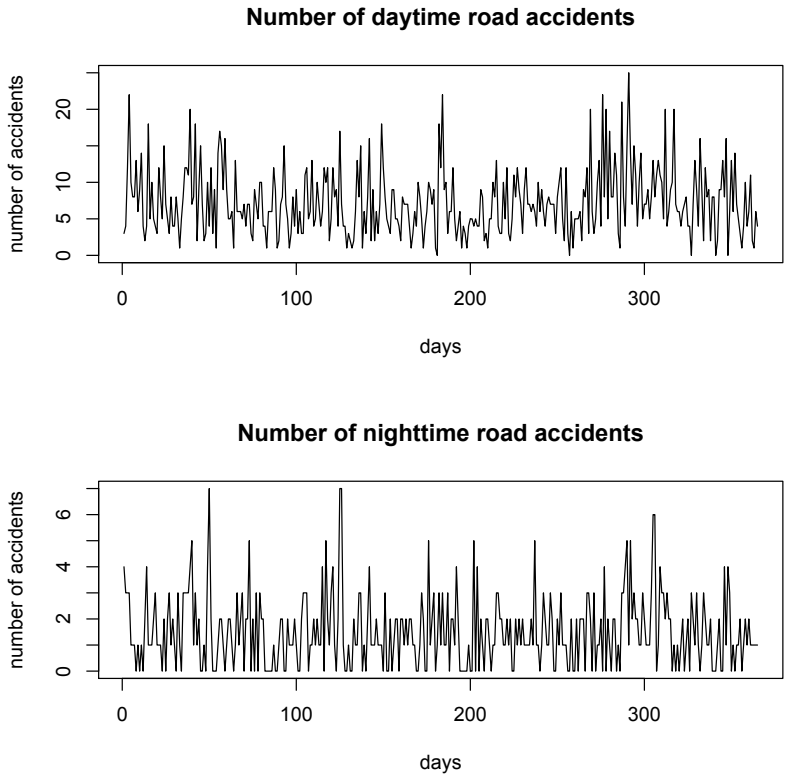

Figure 4.1: Number of daytime and nighttime road accidents in Schiphol area in the Netherlands.

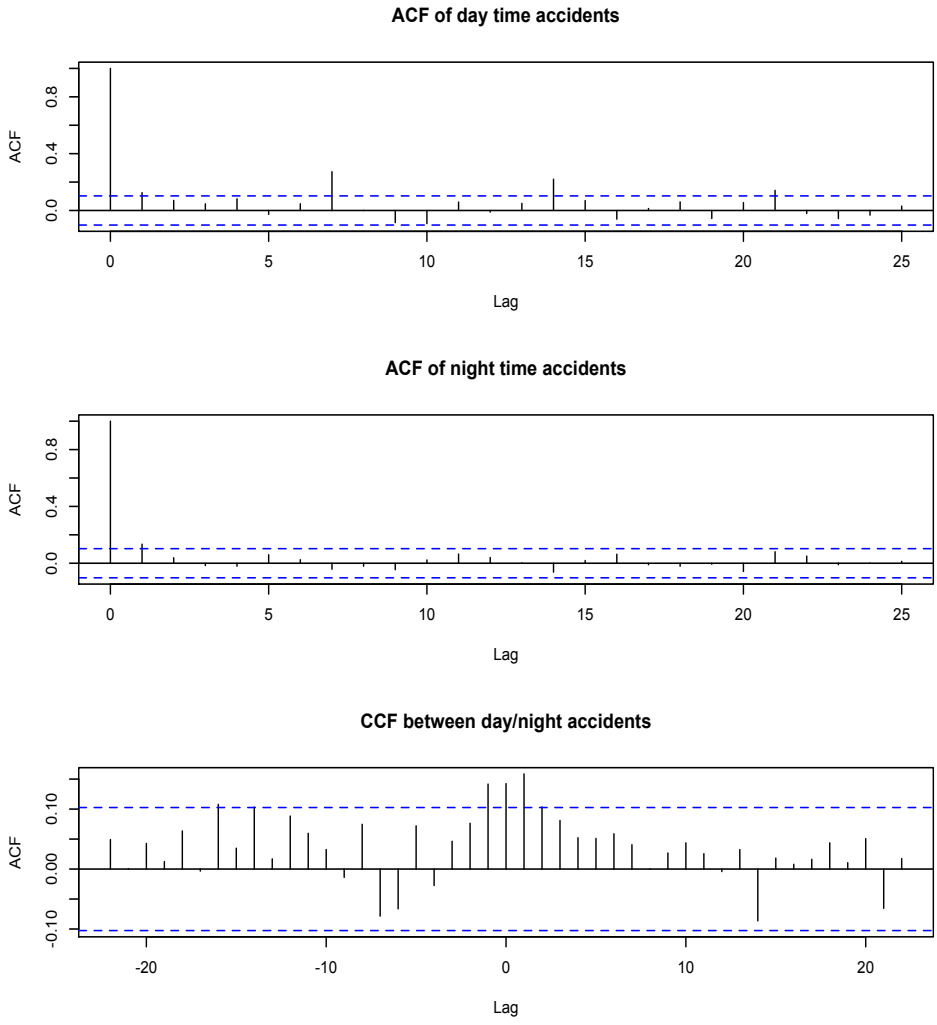

Figure 4.2: Top: autocorrelation of the number of daytime accidents in Schiphol area; Middle: autocorrelation of the number of nighttime accidents; Bottom: cross-correlation between the numbers of daytime and nighttime road accidents.

both time series. It can be seen that the model is capable of capturing the basic structure and fluctuation of each individual time series. In particular, the fitted conditional mean of daytime number of accidents appears to be able to model the seasonal component. Moreover, the residuals appear to be white. For comparison, an INGARCH(7, 7) model with nonzero coefficients only at lags 1 and 7 is fitted to each individual univariate time series. It turns out that modeling jointly increases the log likelihood and reduces AIC and BIC values, for example, the log likelihood and BIC of the BINGARCH are -1682 and 3463, respectively, while those of the INGARCH are -1711 and 3488, respectively. Furthermore, the joint modeling strategy produces smaller prediction scores consistently, for instance, the ranked probability scores (see (2.5.3)) of the time series of daytime accidents of BINGARCH and INGARCH models are 2.39 and 2.46, respectively.

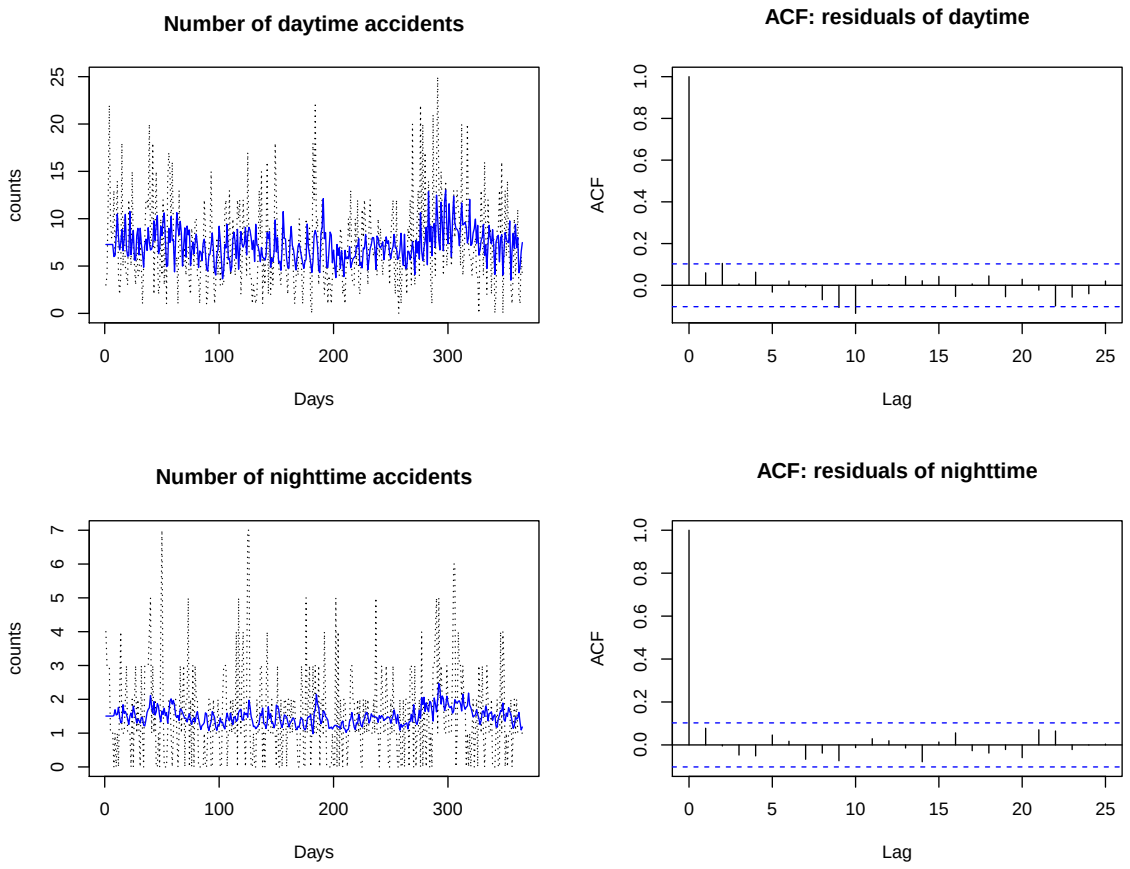

Figure 4.3: Left: fitted conditional mean processes of daytime and nighttime accidents using a BINGARCH model with lag 7 terms; Right: ACF of Pearson residuals of the two time series.

Chapter 5

Conclusions and Future Work

This thesis focuses on developing nonlinear time series models and establishing relevant theory with a view towards applications in which the responses are integer-valued. We first propose a broad class of observation-driven models that is based upon a one-parameter exponential family of distributions and incorporates nonlinear dynamics. This class allows for easy and straightforward estimation of model parameters. The establishment of stability properties of the processes, including the stationarity and ergodicity, is addressed by employing theory from Markov chains and specifically iterated random functions. Several model extensions are considered, including a self-excited threshold Poisson autoregression and the incorporation of explanatory covariates. In addition, given a suitable definition of a multivariate Poisson distribution, a multivariate Poisson autoregression process is proposed to model multivariate time series of counts. As shown in many data applications presented in this thesis, the proposed models are capable of modeling serial dependence in the time series and improving the goodness of fit, especially as compared to some of the main competing observation-driven models in the literature.

Future directions of research:

1. As pointed out in Section 3.2, the self-excited threshold Poisson autoregression is capable of modeling negatively correlated observations. However, the analytical expression for the auto-correlation function is unavailable. It may be worthwhile for the estimation procedure if we can obtain the conditions on the parameters under which the resulting ACF's have negative values.

2. Some further generalizations of the threshold model are desired. For example, the model can have multiple thresholds on lagged observations and the conditional distribution could be extended to a negative binomial distribution and other distributions belonging to the one-parameter exponential family.

3. For the INGARCH model with covariates, the assumptions in Proposition 3.3.1 are due to technical reasons in order to guarantee the φ-irreducibility of the Markov chain. However, it is of interest to study if the conditions could be relaxed, but still yield similar stability properties of the model.

4. In modeling multivariate time series of counts, we only investigate the properties of a bivariate Poisson autoregression. However, can the results be generalized to multivariate Poisson autoregression? How about other multivariate discrete distributions, say, multivariate negative binomial?

5. Even in the case of the bivariate Poisson autoregression, the model is only capable of modeling positive serial dependence between the two time series of counts and the dependence is time-invariant. In future work, we are interested in exploring other possible model formulations to allow for negative or time-varying serial dependence between the two time series.

Chapter 6

Appendix: Markov Chain Theory

6.1 Introduction

This appendix aggregates and provides some useful Markov chain theory, including some preliminary definitions and theorems from Meyn and Tweedie (2009), iterated random functions from Diaconis and Freedman (1999) and Wu and Shao (2004) and τ-weak dependence from Doukhan and Wintenberger (2008) and Dedecker and Prieur (2004), all of which play a key role in all of the results from previous chapters. The proofs will be omitted here and can be found in the indicated references.

6.2 Classical Markov Chain Theory

Definition 6.2.1. A Markov chain $\{X_t\}$ with state space E is *φ-irreducible* if there exists a measure φ on $\mathcal{B}(E)$ such that

$$\sum_{t=1}^{\infty} P^t(x, A) > 0, \quad \text{for all } x \in E,$$

whenever $\varphi(A) > 0$, where $A \in \mathcal{B}(E)$.

Definition 6.2.2. A set $C \in \mathcal{B}(E)$ is called a ν_m-*small set* if there exists an $m > 0$ and a non-trivial measure ν_m on $\mathcal{B}(E)$, such that for all $x \in C$ and $B \in \mathcal{B}(E)$,

$$P^m(x, B) \geq \nu_m(B).$$

Definition 6.2.3. A set $C \in \mathcal{B}(E)$ is called ν_a-*petite* if the chain satisfies

$$\sum_{n=0}^{\infty} P^n(x, B)a(n) \geq \nu_a(B)$$

for all $x \in C$ and $B \in \mathcal{B}(E)$, where ν_a is a non-trivial measure on $\mathcal{B}(E)$.

Definition 6.2.4. A φ-irreducible Markov chain on a general state space is called *strongly aperiodic* if there exists a ν_1-small set A with $\nu_1(A) > 0$.

Definition 6.2.5. A chain is said to be *weak Feller* if its transition probability kernel P maps $C(E)$ to $C(E)$, where

$$P(h(x)) := \int P(x, dy)h(y), \ x \in E,$$

and $C(E)$ represents the class of bounded continuous functions from E to $\mathbb{R}$ and E is the state space of the chain.

Definition 6.2.6. A point $x \in E$ is called *reachable* if for every open set $O \in \mathcal{B}(E)$ containing x,

$$\sum_{n} P^n(y, O) > 0, \text{ for any } y \in E.$$

Definition 6.2.7. A σ-finite measure π is *invariant* if

$$\pi(A) = \int_E \pi(dx)P(x, A), \ A \in \mathcal{B}(E).$$

Definition 6.2.8. A sequence of probabilities $\{\mu_k, k \in \mathbb{Z}_+\}$ is *tight* if for any $\epsilon > 0$, there exists a compact set $C \subset E$ such that $\liminf_{k\to\infty} \mu_k(C) \geq 1 - \epsilon$.

Definition 6.2.9. A chain $\{X_t\}$ is called *bounded in probability on average* if for any initial state $x \in E$, the sequence $\{1/k \sum_{t=1}^{k} P^t(x, \cdot) : k \in \mathbb{Z}_+\}$ is tight.

Theorem 6.2.1. *(Theorem 12.0.1 of Meyn and Tweedie (2009)) If $\{X_t\}$ is a weak Feller chain which is bounded in probability on average, then there exists at least one invariant probability measure.*

Definition 6.2.10. The Markov transition function P is called *equicontinuous* if for any continuous function f with compact support, the sequence of functions $\{P^k f : k \in \mathbb{Z}_+\}$ is equicontinuous on compact sets. A Markov chain which possesses an equicontinuous Markov transition function is called an *e-chain*.

Theorem 6.2.2. *(Theorem 18.8.4 of Meyn and Tweedie (2009)) If $\{X_t\}$ is an e-chain which is bounded in probability on average, then a unique invariant probability measure exists if and only if a reachable state $x^* \in E$ exists.*

Definition 6.2.11. A Markov chain $\{X_t\}$ is *ergodic* if there exists a probability measure π such that

$$\sup_{A \in \mathcal{B}(E)} |P^n(x, A) - \pi(A)| \longrightarrow 0, \quad \text{as } n \to \infty$$

for any $x \in E$, where $P^n(x, A) = P(X_n \in A | x_0 = x)$.

If μ is a signed measure, the total variation norm is

$$\|\mu\|_{TV} = \sup_{f:|f| \leq 1} |\mu(f)| = \sup_A \mu(A) - \inf_A \mu(A).$$

So for an ergodic Markov chain, we have

$$\lim_{n \to \infty} \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} = 2 \lim_{n \to \infty} \sup_A |P^n(x, A) - \pi(A)| = 0.$$

Note that if $\{X_t\}$ is ergodic, then π is its invariant probability measure. To see this, note that

$$\pi(A) = \lim_{n\to\infty} P^{n+1}(x, A) = \lim_{n\to\infty} \int P(y, A)P^n(x, dy) = \int P(y, A)\pi(dy).$$

Definition 6.2.12. A Markov chain $\{X_t\}$ is *geometrically ergodic* if there exists $\rho \in (0, 1)$ such that for any $x \in E$,

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} = o(\rho^n).$$

Remark 6.2.1. A geometrically ergodic Markov chain $\{X_t\}$ starting from its stationary distribution is α-mixing with geometrically decaying rate.

Theorem 6.2.3. *(Theorem in Meyn and Tweedie (2009)) Suppose $\{X_t, t \geq 1\}$ is a Feller chain, and there exist a measure ϕ and a compact set A with $\phi(A) > 0$, such that*

(i) $\{X_t\}$ is ϕ-irreducible,

(ii) there exists a non-negative continuous function $g : E \to \mathbb{R}^1$, such that

$$g(x) \geq 1, \text{ for all } x \in A,$$

and for some $\rho \in (0, 1)$

$$E[g(X_{t+1})|X_t = x] \leq (1 - \rho)g(x), \text{ for all } x \in A^c,$$

then $\{X_t\}$ is geometrically ergodic.

6.3 Iterated Random Functions

Theory from iterated random functions unifies many branches in probability theory. The problem of the existence of stationary distributions and related convergence issues have received extensive study in the literature, for example, Barnsley and Elton (1988), Elton (1990), Arnold (1998), Stenflo (1998), Diaconis and Freedman (1999) and Wu and Shao (2004). Following the notation used by Wu and Shao (2004), we denote (E, d) as a Polish (i.e., complete separate metric) space with Borel sets $\mathcal{B}(E)$. Then an iterated random function system (IRF) on the state space E is defined as

$$X_n = f_{\theta_n}(X_{n-1}), \quad n \in \mathbb{N},$$

where θ and $\{\theta_n, n \in \mathbb{N}\}$ take values in a second measurable space Θ and are independently distributed with identical marginal distribution H. Here $f_\theta(\cdot) = f(\cdot, \theta)$ is the θ-section of a jointly measurable function $f : E \times \Theta \to E$ and X_0 is independent of $\{\theta_n\}_{n\geq 1}$. A Markov chain $\{X_t\}$ can be represented as a system of iterated random functions with carefully chosen $\{\theta_n\}$ such that

$$X_n = f_{\theta_n} \circ f_{\theta_{n-1}} \circ \ldots \circ f_{\theta_1}(x), \tag{6.3.1}$$

which is also known as the *forward process*. To facilitate the investigation, introduce the *backward process* $Z_n(x) = f_{\theta_1} \circ \ldots \circ f_{\theta_n}(x)$. Notice that, for all $x \in E$, $Z_n(x) \stackrel{d}{=} X_n(x)$. So if $Z_n(x)$ converges almost surely to a proper random variable, then $X_n(x)$ converges in distribution.

Definition 6.3.1. (Wu and Shao (2004)) Assume π is an invariant probability measure of the Markov chain $\{X_n\}$. Let X_0 and $X_0' \sim \pi$ be independent of each other and of $\{\theta_n\}_{n\geq 1}$, such that $X_n(X_0')$ can be viewed as a coupled version of $X_n(X_0)$. Then X_n is *geometric moment contracting* if there exist an $\alpha > 0$, a $C = C(\alpha)$ and an

$r = r(\alpha) \in (0,1)$ such that, for all $n \in \mathbb{N}$,

$$\mathrm{E}\{d^\alpha(X_n(X_0), X_n(X_0'))\} \le Cr^n. \tag{6.3.2}$$

Remark 6.3.1. If a Markov chain $\{X_n\}$ is geometric moment contracting, then π is its unique stationary distribution.

The study on IRF revolves around imposing regularity conditions on $f_\theta(\cdot)$ under which the Markov chain enjoys stationarity and some mixing conditions. In our research, the following two conditions are usually verified for the models under study, see Wu and Shao (2004).

Condition 1. There exists a $y_0 \in E$ and an $\alpha > 0$ such that

$$I(\alpha, y_0) := \mathrm{E}\{d^\alpha(y_0, f_\theta(y_0))\} = \int_\Theta d^\alpha(y_0, f_\theta(y_0)) H(d\theta) < \infty. \tag{6.3.3}$$

Condition 2. There exist an $x_0 \in E$, an $\alpha > 0$, an $r(\alpha) \in (0,1)$ and a $C(\alpha) > 0$ such that

$$\mathrm{E}\{d^\alpha(X_n(x), X_n(x_0))\} \le C(\alpha) r^n(\alpha) d^\alpha(x, x_0) \tag{6.3.4}$$

for all $x \in E$ and $n \in \mathbb{N}$.

Definition 6.3.2. A random variable is said have an *algebraic tail* is there exist $A, B > 0$ such that $P(|Y| > y) < A/y^B$ for all $y > 0$.

Theorem 6.3.1. *(Diaconis and Freedman (1999)) Assume that Condition 1 holds, that*

$$E[\log K_\theta] = \int_\Theta \log K_\theta H(d\theta) < 0, \quad \textit{where } K_\theta = \sup_{x \neq x'} \frac{d(f_\theta(x), f_\theta(x'))}{d(x, x')},$$

and that K_θ has an algebraic tail. Then there exists a unique stationary distribution π for (6.3.1) and $Z_n(x) \to Z_\infty \sim \pi$ at a geometric rate. The limit Z_∞ does not depend on x.

Theorem 6.3.2. *(Wu and Shao (2004)) Suppose that Conditions 1 and 2 hold. Then there exists a random variable Z_∞ such that for all $x \in E$, $Z_n(x) \to Z_\infty$ almost surely. The limit Z_∞ is $\sigma\{\theta_1, \theta_2, \ldots\}$-measurable and does not depend on x. Moreover, for every $n \in \mathbb{N}$,*

$$E\{d(Z_n(x), Z_\infty)^\alpha\} \leq Cr(\alpha)^n,$$

where $C > 0$ depends solely on x, x_0, y_0 and α, and $0 < r(\alpha) < 1$. In addition, (6.3.2) holds.

6.4 Weak Dependence

The concept of τ-weak dependence, which is less restrictive than mixing conditions (see Andrews (1984) for an example), is used in this thesis. Readers can refer to Doukhan and Wintenberger (2008) and Dedecker and Prieur (2004) for details. To better understand τ-weak dependence, we first give the definitions of β- and τ-coefficients.

Definition 6.4.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{M}$ be a σ-algebra of $\mathcal{A}$,X be a E-valued random variable with distribution P_X, and $P_{X|\mathcal{M}}$ be a conditional distribution of X given $\mathcal{M}$. The β-mixing coefficient between $\mathcal{M}$ and $\sigma(X)$ can be defined as

$$\beta(\mathcal{M}, \sigma(X)) = \frac{1}{2}\|V(P_{X|\mathcal{M}})\|_1, \tag{6.4.1}$$

where

$$V(P_{X|\mathcal{M}}) = \sup\Big\{\Big|\int f(x)P_{X|\mathcal{M}}(dx) - \int f(x)P_X(dx)\Big| : \|f\|_\infty \leq 1\Big\}. \tag{6.4.2}$$

One of the most important properties of β-mixing coefficient is Berbee's coupling lemma (Berbee (1979)): if Ω is rich enough, then there exists a random variable X^* independent of $\mathcal{M}$ and distributed as X such that $\mathbb{P}(X \neq X') = \beta(\mathcal{M}, \sigma(X))$.

In the sequel, denote $\|\cdot\|_m$ as the usual L^m-norm of the E-valued random variable X defined on Ω, i.e., $\|X\|_m^m = \mathrm{E}\|X\|^m$ for $m \geq 1$. For the function $h : E \to \mathbb{R}$, denote

$$\mathrm{Lip}(h) = \sup_{x \neq y} \frac{|h(x) - h(y)|}{\|x - y\|}.$$

The space $\Lambda_1(E)$ is the set of functions $h : E \to \mathbb{R}$ such that $\mathrm{Lip}(h) \leq 1$. The concept of τ-coefficient relies on the set $\Lambda_1(E)$.

Definition 6.4.2. If the E-valued random variable X is integrable, i.e., $\|X\|_1 < \infty$, then the τ-coefficient between $\mathcal{M}$ and $\sigma(X)$ is defined as

$$\tau(\mathcal{M}, X) = \|W(P_{X|\mathcal{M}})\|_1, \tag{6.4.3}$$

where

$$W(P_{X|\mathcal{M}}) = \sup\Big\{\Big| \int f(x) P_{X|\mathcal{M}}(dx) - \int f(x) P_X(dx)\Big| : f \in \Lambda_1(E)\Big\}. \tag{6.4.4}$$

The coupling also works for the τ-weak dependence: if Ω is rich enough, the coefficient $\tau(\mathcal{M}, X)$ is the infimum of $\|X - Y\|_1$, where Y is independent of $\mathcal{M}$ and distributed as X, and this infimum can be reached by some particular random variable X^* (see e.g., Major (1978)).

Doukhan and Wintenberger (2008) considers the τ-weak dependence structure of a *chain with infinite memory*. Here we present their results tailored to Markov chains (6.3.1). Using Definition 6.4.2, the dependence between the past of the of the $\{X_t\}_{t \in \mathbb{Z}}$ and its future k-tuple can be assessed: consider the norm $\|x - y\| =$

$\|x_1 - y_1\| + \ldots + \|x_k - y_k\|$ on E^k, set $\mathcal{M}_p = \sigma(X_t, t \geq p)$ and define

$$\tau_k(r) = \max_{1 \leq l \leq k} \frac{1}{l} \sup\big\{\tau(\mathcal{M}_p, (X_{j_1}, X_{j_2}, \ldots, X_{j_l})) : p + r \leq j_1 < \ldots < j_l\big\},$$
$$\tau_\infty(r) = \sup_{k>0} \tau_k(r).$$

For the sake of simplicity, $\tau_\infty(r)$ is denoted as $\tau(r)$. Finally, the time series $\{X_t\}_{t\in\mathbb{Z}}$ is *τ-weakly dependent* when its coefficients $\tau(r)$ tend to 0 as r tends to infinity.

Theorem 6.4.1. *(Doukhan and Wintenberger (2008)) For the Markov chain (6.3.1), if for all $x, y \in E$,*

$$E\|f(x,\theta) - f(y,\theta)\| \leq a\|x - y\|,$$

where $a \in (0,1)$ and $\mu_1 = E\|f(0,\theta)\|_1 < \infty$, then there exists a τ-weakly dependent stationary solution $\{X_t\}$ such that $E\|X_0\| < \infty$ and $\tau(r) \leq 2\mu_1(1-a)^{-1}a^r$ for $r \geq 1$. In addition, $\{X_t\}$ is the unique causal Bernoulli shift solution and is automatically an ergodic process.

Bibliography

Andrews, D. (1984) Nonstrong mixing autoregressive processes. *Journal of Applied Probability*, **21(4)**, 930–934.

Arnold, L. (1998) *Random Dynamical Systems.* Springer, Berlin.

Barnsley, M. F. and Elton, J. H. (1988) A new class of Markov processes for image coding. *Adv. Appl.Prob.*, **20**, 14–32.

Berbee, H. (1979) *Random walks with stationary increments and renewal theory.* Math. Cent. Tracts.

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

Billingsley, P. (1999) *Convergence of probability measures. (2nd edition).* New York: Wiley.

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

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307327.

Brijs, T., Karlis, D. and Wets, G. (2008) Studying the effect of weather conditions on daily crash counts using a discrete time series model. *Accident Analysis and Prevention*, **40**, 1180–1190.

Brockwell, A. E. (2007) Universal residuals: A multivariate transformation. *Statistics and Probability Letters*, **77(14)**, 1473–1478.

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

Campbell, M. J. (1994) Time series regression for counts: an investigation into the relationship between sudden infant death syndrome and environmental temperature. *J. R. Statist. Soc. A*, **157**, 191–208.

Chan, K. and Ledolter, J. (1995) Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, **90**, 242–252.

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. *Scandinavian Journal of Statistics*, **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**, 491506.

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

Davis, R. and Rodriguez-Yam, G. A. (2005) Estimation for state-space models: an approximate likelihood approach. *Statistica Sinica*, **15**, 381–406.

Davis, R. and Yao, C.-Y. (2009) Comments on pairwise likelihood in time series models. *Statistica Sinica*, **21**, 255–277.

Dedecker, J. and Prieur, C. (2004) Coupling for τ-dependent sequences and applications. *Journal of Theoretical Probability*, **17(4)**, 861–855.

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

Doukhan, P. (1994) *Mixing: Properties and Examples. Lecture notes in Statistics 85.* Springer-Verlag.

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

Doukhan, P. and Wintenberger, O. (2008) Weakly dependent chains with infinite memory. *Stochastic Processes and their Applications*, **118(11)**, 19972013.

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

Durbin, J. and Koopman, S. (2001) *Time Series Analysis by State Space Methods (Oxford Statistical Science Series)*. Oxford University Press.

Elton, J. H. (1990) A multiplicative ergodic theorem for Lipschitz maps. *Stochastic Processes and their Applications*, **34**, 39–47.

Ferland, R., Latour, A. and Oraichi, D. (2006) Integer-valued GARCH process. *Journal of Time Series Analysis*, **27(6)**, 923–942.

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

Gelfand, I. (1941) Normierte ringe. *Rec. Math. [Mat. Sbornik] N.S.*, **9(51)**, 324.

Guttorp, P. (1991) *Statistical Inference for Branching Processes*. New York: John Willey & Sons.

Hardy, E. (1996) *Modélisation de type ARCH pour séries chronologiques à valeurs entières.* Master's thesis, Départment de Mathématiques, Université du Québec à Montréal.

Heinen, A. and Rengifo, E. (2003) Multivariate modelling of time series count data: an autoregressive conditional Poisson model. core discussion paper 25. Tech. rep., Catholic University of Louvain.

Johnson, N., Kotz, S. and Balakrishnan, N. (1997) *Multivariate discrete distributions*. New York: Wiley.

Jung, R. and Tremayne, A. (2011) Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, **95**, 59–91.

Kitagawa, G. (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, **5(1)**, 1–25.

Kocherlakota, S. and Kocherlakota, K. (1992) *Bivariate discrete distributions, statistics: textbooks and monographs, Vol. 132*. New York: Markel Dekker.

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

Levine, N., Kim, K. and Nitz, L. (1995a) Daily fluctuations in Honolulu motor vehicle accidents. *Accident Analysis and Prevention*, **27**, 785–796.

Levine, N., Kim, K. and Nitz, L. (1995b) Spatial analysis of Honolulu motor vehicle crashes: I. spatial patterns. *Accident Analysis and Prevention*, **27**, 663–674.

Major, P. (1978) On the invariance principle for sums of identically distributed random variables. *Journal of Multivariate analysis*, **8**, 487–517.

Meyn, S. and Tweedie, R. (2009) *Markov Chains and Stochastic Stability (2nd edition).* Cambridge University Press.

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

Pedeli, X. and Karlis, D. (2010) On composite likelihood estimation of a multivariate Poisson INAR(1) models. In *Proceedings of the 25th Workshop in statistical Modelling, July 2010, Glaskow, pp 429-432.*

Pedeli, X. and Karlis, D. (2011) A bivariate INAR(1) process with application. *Statistical modelling*, **11**, 325–349.

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 Series in Statistical and Probabilistic Mathematics).* Cambridge University Press.

Samia, N. and Chan, K. (2010) Maximum likelihood estimation of a generalized threshold stochastic regression model. *Biometrika*, **98 (2)**, 433–448.

Stenflo, Ö. (1998) *Ergodic theorems for iterated function systems controlled by stochastic sequences.* Ph.D. thesis, Umeå University.

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.* New York: Oxford University Press.

Wang, W., Liu, H., Davis, R., Yao, J. F. and Li, W. K. (2012) Self-excited Threshold Poisson Autoregression. Tech. rep., Columbia University, The University of Hong Kong.

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

Yu, Y. (2009) Stochastic ordering of exponential family distributions and their mixtures. *Journal of Applied Probability*, **46**, 244–254.

Zeger, S. (1988) A regression model for time series of counts. *Biometrika*, **75**, 6219.