Stationarity of Generalized Autoregressive Moving Average Models

Dawn B. Woodard David S. Matteson
Shane G. Henderson
School of Operations Research and Information Engineering
Cornell University *

March 30, 2011

Abstract

Time series models are often constructed by combining nonstationary effects such as trends with stochastic processes that are believed to be stationary. Although stationarity of the underlying process is typically crucial to ensure desirable properties or even validity of statistical estimators, there are numerous time series models for which this stationarity is not yet proven. A major barrier is that the most commonly-used methods assume φ-irreducibility, a condition that can be violated for the important class of discrete-valued observation-driven models.

We show (strict) stationarity for the class of Generalized Autoregressive Moving Average (GARMA) models, which provides a flexible analogue of ARMA models for count, binary, or other discrete-valued data. We do this from two perspectives. First, we show stationarity and ergodicity of a perturbed version of the GARMA model, and show that the perturbed model yields parameter estimates that are arbitrarily close to those of the original model. This approach utilizes the fact that the perturbed model is φ-irreducible. Second, we show that the original GARMA model has a unique stationary distribution (so is strictly stationary when initialized in that distribution).

Keywords: stationary; time series; ergodic; Feller; drift conditions; irreducibility.

1 Introduction

Stationarity is a fundamental concept in time series modeling, capturing the idea that the future is expected to behave like the past; this assumption is inherent in any attempt to forecast the future. Many time series models are created by combining nonstationary effects such as

*The authors thank the referees for their helpful suggestions that led to many improvements in this article. They also thank Prof. Natesh Pillai for several helpful conversations. This work was supported in part by National Science Foundation Grant Number CMMI-0926814. Portions of this work were completed while Shane Henderson was visiting the Mathematical Sciences Institute at the Australian National University and he thanks them for their hospitality and support.

trends, covariate effects, and seasonality with a stochastic process that is known or believed to be stationary. Alternatively, they can be defined by partial sums or other transformations of a stationary process. The properties of statistical estimators for particular models are then established via the relationship to the stationary process; this includes consistency of parameter estimators and of standard error estimators (Brockwell and Davis 1991, Chap. 7-8).

However, (strict) stationarity can be nontrivial to establish, and many time series models currently in use are based on processes for which it has not been proven. Strict stationarity (henceforth, "stationarity") of a stochastic process $\{X_n\}_{n\in\mathbb{Z}}$ means that the distribution of the random vector $(X_n, X_{n+1} \ldots, X_{n+j})$ does not depend on n, for any $j \geq 0$ (Billingsley 1995, p.494). Sometimes weak stationarity (constant, finite first and second moments of the process $\{X_n\}_{n\in\mathbb{Z}}$) is proven instead, or simulations are used to argue for stationarity.

The most common approach to establishing strict stationarity and ergodicity (defined as in Billingsley 1995, p.494) is via application of Lyapunov function methods (also known as drift conditions) to a Markov chain that is related to the time series model. However, Lyapunov function methods are almost always used in conjunction with an assumption of φ-irreducibility, which can be violated by discrete-valued observation-driven time series models. Such models are important since (due to the simplicity of evaluating the likelihood function) they are typically the best option for modeling very long count- or binary-valued time series.

We address this challenge to show for the first time stationarity of several models designed for discrete-valued time series, including the class of Generalized Autoregressive Moving Average (GARMA) models (Benjamin et al., 2003). We do this from two perspectives. First, we show stationarity and ergodicity of a perturbed version of the model. Such ergodicity results for a perturbed model can be used in some cases to show asymptotic normality of parameter estimators for the original model (Fokianos et al., 2009; Fokianos and Tjostheim, 2011, 2010). We show that the perturbed and original models are closely related in the sense that the perturbed model yields (finite-sample) parameter estimates that are arbitrarily close to those from the original model. This result is given under weak conditions that encompass nearly all discrete-valued time series models, and utilizes the fact that the perturbed model is φ-irreducible. It implies that a researcher can choose to use the perturbed model without substantially affecting (finite-sample) parameter estimates, in order to get strong theoretical properties associated with stationarity and ergodicity, such as consistent estimation of the mean and lagged covariances of the process, and more generally the expectation of any integrable function (Billingsley 1995, p.495). However, our result does not immediately imply stationarity properties for the original model. Indeed, it is not clear that one can always use such perturbation results to show asymptotic normality of estimators in the original model, and when it is possible model-specific approaches may be needed.

Instead of pursuing such a model-specific approach we maintain attention on a broad class of models and investigate a second approach. We show that the original models have a unique stationary distribution (so are strictly stationary when initialized in that distribution), by using Feller properties of the chain. To our knowledge Feller properties have not previously been used to show uniqueness of the stationary distribution for discrete-valued time series models (although they have been used to show existence of a stationary distribution; Davis, Dunsmuir, and Streett 2003). This approach is very general and has the potential to be used for showing that wide classes of discrete-valued models have unique stationary distributions.

Our stationarity results for the original models potentially form the foundation for showing consistency and asymptotic normality of parameter estimates in the GARMA model class in its general form. Unlike the results for the perturbed model, consistent estimation of the expectation of integrable functions is not immediate from our stationarity results for the

unperturbed model; we leave this to future work.

GARMA models generalize autoregressive moving average models to exponential-family distributions, naturally handling count- and binary-valued data among others. They can also be seen as an extension of generalized linear models to time series data. The numerous applications of these models include predicting numbers of births (Léon and Tsai, 1998), modeling poliomyelitis cases (Benjamin et al., 2003), and predicting valley fever incidence (Talamantes et al., 2007). The main stationarity result that currently exists for GARMA models is weak stationarity in the case of an identity link function; unfortunately this excludes the most popular of the count-valued models (Benjamin et al., 2003). Zeger and Qaqish (1988) have also used a connection to branching processes to show stationarity for a special case of the purely autoregressive Poisson log-link GARMA model. The stationarity of particular models related to Poisson GARMA has also been addressed by Davis et al. (2003) (log link case) and Ferland et al. (2006) (linear link case).

In Section 2 we describe Lyapunov function methods, give our justification for using these methods on perturbed discrete-valued time series models, and give background on the use of Feller properties to show stationarity. In Section 3 we illustrate the perturbation approach by applying it to a specific count-valued threshold model, and in Section 4 we use the perturbation method to prove stationarity for the class of perturbed GARMA models. In Section 5 we show stationarity of the original GARMA and count-valued threshold models using Feller properties.

2 Stationarity of Observation-Driven Models

For a real-valued process $\{Y_n\}_{n\in\mathbb{N}}$, let $Y_{n:m} = (Y_n, Y_{n+1}, \ldots, Y_m)$ where $n \leq m$. An observation-driven time series model for $\{Y_n\}_{n\in\mathbb{N}}$ has the form:

$$Y_n | Y_{0:n-1} \sim \psi_\nu(\cdot; \mu_n) \tag{1}$$

$$\mu_n = h_{\theta,n}(Y_{0:n-1}) \tag{2}$$

for functions $h_{\theta,n}$ parameterized by θ and some density function ψ_ν (typically with respect to counting or Lebesgue measure) that can depend on both time-invariant parameters ν and the time-dependent quantities μ_n (Zeger and Qaqish, 1988; Davis et al., 2003; Ferland et al., 2006). Observation-driven models are desirable because the likelihood function for the parameter vector (θ, ν) can be evaluated explicitly. The alternative class of parameter-driven models (Cox, 1981; Zeger, 1988), by contrast, incorporates latent random innovations which typically make explicit evaluation of the likelihood function impossible, so that one must resort to approximate inference or computationally intensive Monte Carlo integration over the latent process (Chan and Ledolter, 1995; Durbin and Koopman, 2000; Jung et al., 2006). These methods do not scale well to very long time series, so observation-driven models are typically the better option in this case.

Observation-driven models are usually constructed via a Markov-p structure for μ_n, meaning that for $n \geq p$

$$\mu_n = g_\theta(Y_{n-p:n-1}, \mu_{n-p:n-1}) \tag{3}$$

for some function g_θ and for fixed initial values $\mu_{0:p-1}$. This structure implies that the vector $\mu_{n-p:n-1}$ forms the state of a Markov chain indexed by n. In this case it is sometimes possible to prove stationarity and ergodicity of $\{Y_n\}_{n\in\mathbb{N}}$ by first showing these properties for the multivariate Markov chain $\{\mu_{n-p:n-1}\}_{n\geq p}$, then "lifting" the results back to the time series model

$\{Y_n\}_{n\in\mathbb{N}}$. In particular, showing that $\{\mu_{n-p:n-1}\}_{n\geq p}$ is φ-irreducible, aperiodic and positive Harris recurrent (defined below) implies that it has a unique stationary distribution π, and that if $\mu_{0:p-1} \sim \pi$ then $\{\mu_{n-p:n-1}\}_{n\geq p}$ is a stationary and ergodic process. That $\{Y_n\}_{n\in\mathbb{N}}$ is also stationary and ergodic is seen as follows. Conditional on $\{\mu_n\}_{n\in\mathbb{N}}$, the Y_n are independent across n and each Y_n has a distribution that is a function of only $\mu_{n:n+p}$ (since $Y_n \sim \psi_\nu(\mu_n)$ and since the values $\mu_{n+1:n+p}$ depend on Y_n). Therefore there is a deterministic function f such that one can simulate $\{Y_n\}$ conditional on $\{\mu_n\}$ by: (a) generating an i.i.d. sequence of Uniform$(0,1)$ random variables U_n, and (b) setting $Y_n = f(\mu_{n:n+p}, U_n)$. The multivariate process $\{(\mu_{n-p:n-1}, U_n)\}_{n\geq p}$ is stationary and ergodic, and so Thm. 36.4 of Billingsley (1995) shows that its transformation $\{Y_n\}$ is also stationary and ergodic.

2.1 Stationarity of a Perturbed Process

We will give an approach based on a perturbed version of the discrete-valued model, and justify its use. First we describe the use of drift conditions to show stationarity and ergodicity of φ-irreducible processes. For a general Markov chain $X = \{X_n\}_{n\in\mathbb{N}}$ on state space S with σ-algebra $\mathcal{F}$ define $T^n(x, A) = \Pr(X_n \in A | X_0 = x)$ for $A \in \mathcal{F}$ to be the n-step transition probability starting from state $X_0 = x$. The appropriate notion of irreducibility when dealing with a general state space is that of φ-irreducibility, since general state space Markov chains may never visit the same point twice.

Definition 1. *A Markov chain X is* φ-irreducible *if there exists a nontrivial measure φ on $\mathcal{F}$ such that, whenever $\varphi(A) > 0$, $T^n(x, A) > 0$ for some $n = n(x, A) \geq 1$, for all $x \in S$.*

The notion of aperiodicity in general state space chains is the same as that seen in countable state space chains, namely that one cannot decompose the state space into a finite partition of sets where the chain moves successively from one set to the next in sequence, with probability one. For a more precise definition, see Meyn and Tweedie (1993), Sec. 5.4.

We need one more definition before we can present drift conditions.

Definition 2. *A set $A \in \mathcal{F}$ is called a* small *set if there exists an $m \geq 1$, a nontrivial measure ν on $\mathcal{F}$, and a $\lambda > 0$ such that for all $x \in A$ and all $C \in \mathcal{F}$, $T^m(x, C) \geq \lambda\nu(C)$.*

Small sets are a fundamental tool in the analysis of general state space Markov chains because, among other things, they allow one to apply regenerative arguments to the analysis of a chain's long-run behavior. Regenerative theory is indeed the fundamental tool behind the following result, which is a special case of Theorem 14.0.1 in Meyn and Tweedie (1993). Let $E_x(\cdot)$ denote the expectation under the probability $P_x(\cdot)$ induced on the path space of the chain when the initial state X_0 is deterministically x.

Theorem 1. ***(Drift Conditions):*** *Suppose that $X = \{X_n\}_{n\in\mathbb{N}}$ is φ-irreducible on S. Let $A \subset S$ be small, and suppose that there exist $b \in (0, \infty)$, $\epsilon > 0$, and a function $V : S \to [0, \infty)$ such that for all $x \in S$,*

$$E_x V(X_1) \leq V(x) - \epsilon + b\mathbf{1}_{\{x\in A\}}. \tag{4}$$

Then X is positive Harris recurrent.

The function V is called a Lyapunov function or energy function. The condition (4) is known as a drift condition, in that for $x \notin A$, the expected energy V drifts towards zero by at least ϵ. The indicator function in (4) asserts that from a state $x \in A$, any energy increase is bounded (in expectation).

Positive Harris recurrent chains possess a unique stationary probability distribution π. If X_0 is distributed according to π then the chain X is a stationary process. If the chain is also aperiodic then X is ergodic, in which case if the chain is initialized according to some other distribution, then the distribution of X_n will converge to π as $n \to \infty$.

Hence, the drift condition (4), together with aperiodicity, establishes ergodicity. A stronger form of ergodicity, called geometric ergodicity, arises if (4) is replaced by the condition

$$E_x V(X_1) \leq \beta V(x) + b\mathbf{1}_{\{x \in A\}} \tag{5}$$

for some $\beta \in (0,1)$ and some $V : S \to [1, \infty)$ (note the change in the range of V). Indeed, (5) implies (4). Either of these criteria are sufficient for our purposes.

A problem can occur, however, when we attempt to apply this method for proving stationarity to an observation-driven time series model given by (1) and (3): the Markov chain $\{\mu_{n-p:n-1}\}_{n \geq p}$ may not be φ-irreducible. This occurs, for instance, whenever Y_n can only take a countable set of values and the state space of $\mu_{n-p:n-1}$ is $\mathbb{R}^p$. Then, given a particular initial vector $\mu_{0:p-1}$ the set of possible values for μ_n is countable. In fact, for any fixed initialization $\mu_{0:p-1}$ there is a countable set $A \subset \mathbb{R}^p$ such that $\sum_{n=p}^{\infty} \Pr(\mu_{n-p:n-1} \in A | \mu_{0:p-1}) = 1$, and distinct initial vectors $\mu_{0:p-1}$ can have distinct sets A. For a simpler example of a Markov chain with the same property, consider the stochastic recursion defined by $X_n = [X_{n-1} + Y_n] \mod 1$ where $\{Y_n\}_{n \geq 1}$ are i.i.d. discrete random variables on the rationals and $x \mod 1$ is the fractional part of x. If X_0 is rational, then so is X_n for all $n \geq 1$, while if X_0 is irrational then so is X_n for all $n \geq 1$. Also, the set of states that can be reached from any fixed X_0 is countable.

However, by adding small real-valued perturbations to a discrete-valued time series model one can obtain a φ-irreducible process. We do this by returning to the most general framework (1) and (2), and replacing $h_{\theta,n}$ with a function of two inputs:

$$\begin{aligned} Y_n^{(\sigma)} | Y_{0:n-1}^{(\sigma)} &\sim \psi_\nu(\cdot; \mu_n^{(\sigma)}) \\ \mu_n^{(\sigma)} &= h_{\theta,n}(Y_{0:n-1}^{(\sigma)}, \sigma Z_{0:n-1}) \end{aligned} \tag{6}$$

where the $Z_i \overset{\text{iid}}{\sim} \phi$ are random perturbations having density function ϕ (typically with respect to Lebesgue measure), $\sigma > 0$ is a scale factor associated with the perturbation, and $h_{\theta,n}(\cdot, \sigma Z_{0:n-1})$ is a continuous function of $Z_{0:n-1}$ such that $h_{\theta,n}(y, 0) = h_{\theta,n}(y)$ for any y. The value $\mu_0^{(\sigma)}$ is a fixed constant that we take to be independent of σ, so that $\mu_0^{(\sigma)} = \mu_0$. When the perturbed model is constructed to be φ-irreducible, one can then apply drift conditions to prove its stationarity.

We will show that using the perturbed instead of the original model has an arbitrarily small effect on the (finite-sample) parameter estimates. We do this by proving that the likelihood of the parameter vector $\eta = (\theta, \nu)$ calculated using (6) converges uniformly to the likelihood calculated using (2) as $\sigma \to 0$. More precisely, the joint density of the observations $Y = Y_{0:n}^{(\sigma)}$ and first n perturbations $Z = Z_{0:n-1}$, conditional on the parameter vector η, the perturbation scale σ, and the initial value μ_0, is

$$\begin{aligned} f(Y, Z | \eta, \sigma, \mu_0) &= f(Z | \eta, \sigma, \mu_0) \times f(Y | Z, \eta, \sigma, \mu_0) \\ &= \left[\prod_{k=0}^{n-1} \phi(Z_k) \right] \prod_{k=0}^{n} \psi_\nu(Y_k^{(\sigma)}; \mu_k(\sigma Z)) \end{aligned}$$

where $\mu_k(\sigma Z)$ is the value of $\mu_k^{(\sigma)}$ induced by the perturbation vector σZ through (6), with $\mu_0(\sigma Z) = \mu_0$. The likelihood function for the parameter vector η implied by the perturbed

model is the marginal density of Y integrating over Z, i.e.,

$$\mathcal{L}_\sigma(\eta) = f(Y|\eta, \sigma, \mu_0) = \int f(Y, Z|\eta, \sigma, \mu_0)\, dZ.$$

Here we have placed a subscript σ on the likelihood function to emphasize its dependence on σ. Let the likelihood function without the perturbations be denoted by $\mathcal{L}$, so that

$$\mathcal{L}(\eta) = \prod_{k=0}^{n} \psi_\nu(Y_k; \mu_k(0)).$$

Theorem 2. *Under regularity conditions (a) & (b) below, the likelihood function $\mathcal{L}_\sigma$ based on the perturbed model (6) converges uniformly on any compact set K to the likelihood function $\mathcal{L}$ based on the original model (2), i.e.,*

$$\sup_{\eta \in K} |\mathcal{L}_\sigma(\eta) - \mathcal{L}(\eta)| \stackrel{\sigma \to 0}{\longrightarrow} 0$$

for any fixed sequence of observations $y_{0:n}$ and conditional on the initial value μ_0. So if $\mathcal{L}$ is continuous in η and has a finite number of local maxima and a unique global maximum on K, the maximum-likelihood estimate of η based on $\mathcal{L}_\sigma$ converges to that based on $\mathcal{L}$. Also, Bayesian inferences based on $\mathcal{L}_\sigma$ converge to those based on $\mathcal{L}$, in the sense that the posterior probability of any measurable set A using likelihood $\mathcal{L}_\sigma$ (and restricting to a compact set) converges to that using $\mathcal{L}$.

Regularity Conditions:

(a) For any fixed y the function $\psi_\nu(y; \mu)$ is bounded and Lipschitz continuous in μ, uniformly in $\eta \in K$.

(b) For each n, $\mu_n(\sigma Z)$ is Lipschitz in some bounded neighborhood of zero, uniformly in $\eta \in K$.

Assumption (a) holds, e.g., for $\psi_\nu(y; \mu)$ equal to a Poisson or binomial density with mean μ, or a negative binomial density with mean μ and precision parameter ν. As we will see for several models, $\mu_n(\sigma Z)$ can easily be constructed to satisfy (b). Theorem 2 is proven in Appendix A.1.

Theorem 2 says that one can choose to use the perturbed model (with fixed and sufficiently small perturbation scale σ) instead of the original model, without significantly affecting finite-sample parameter estimates, in order to get the strong theoretical properties associated with stationarity and ergodicity. These include consistent estimation of the mean and lagged covariances of the process. Although we have shown that the perturbed and original models are closely related, and although one can use drift conditions to show stationarity and ergodicity properties of the perturbed model, this approach does not yield stationarity and ergodicity properties for the original model. This is due to the substantial technical difficulty associated with interchanging the limits $\sigma \to 0$ and $n \to \infty$. Theorem 2 addresses the case of a fixed number of observations n, as $\sigma \to 0$, while consistency of parameter estimation for the perturbed model is a statement about $n \to \infty$ for fixed σ. Due to this limitation of the perturbation approach we separately address stationarity of the original process, as outlined in Section 2.2.

2.2 Stationarity of the Original Process

When the chain $\{\mu_{n-p:n-1}\}_{n \geq p}$ associated with the observation-driven time-series model is not φ-irreducible we will see that one can instead use Feller properties to prove that it has a

unique stationary distribution. We address existence of a stationarity distribution first, then uniqueness of that distribution.

In the absence of φ-irreducibility, the "weak Feller" condition can be combined with a drift condition to show existence of a stationary distribution. A chain evolving on a complete separable metric space S is said to be "weak Feller" if the transition kernel $T(x,\cdot)$ satisfies $T(x,\cdot) \Rightarrow T(y,\cdot)$ as $x \to y$, for any $y \in S$ and where $\Rightarrow$ indicates convergence in distribution; see, e.g., Section 6.1.1 of Meyn and Tweedie (1993) and Theorem 25.8 (i) and (ii) of Billingsley (1995).

Theorem 3. *(Tweedie, 1988) Suppose that S is a locally compact complete separable metric space with $\mathcal{F}$ the Borel σ-field on S, and that the Markov chain $\{X_n\}_{n\in\mathbb{N}}$ with transition kernel T is weak Feller. Let $A \in \mathcal{F}$ be compact, and suppose that there exist $b \in (0,\infty)$, $\epsilon > 0$, and a function $V : S \to [0,\infty)$ such that for all $x \in S$ the drift condition (4) holds. Then there exists a stationary distribution for T.*

Uniqueness of the stationary distribution can be established using the "asymptotic strong Feller" property, defined as follows (Hairer and Mattingly, 2006). Let S be a Polish (complete, separable, metrizable) space. A "totally separating system of metrics $\{d_n\}_{n\in\mathbb{N}}$ for S" is a set of metrics such that for any $x, y \in S$ with $x \neq y$, the value $d_n(x,y)$ is nondecreasing in n and $\lim_{n\to\infty} d_n(x,y) = 1$. A metric d on S implies the following distance between probability measures μ_1 and μ_2:

$$\|\mu_1 - \mu_2\|_d = \sup_{\mathrm{Lip}_d\phi=1} \left(\int \phi(x)\mu_1(dx) - \int \phi(x)\mu_2(dx) \right) \tag{7}$$

where

$$\mathrm{Lip}_d\phi = \sup_{x,y\in S: x\neq y} \frac{|\phi(x) - \phi(y)|}{d(x,y)}$$

is the minimal Lipschitz constant for ϕ with respect to d. Using these definitions, a chain is asymptotically strong Feller if, for every fixed $x \in S$, there is a totally separating system of metrics $\{d_n\}$ for S and a sequence $t_n > 0$ such that

$$\lim_{\gamma\to 0} \limsup_{n\to\infty} \sup_{y\in B(x,\gamma)} \|T^{t_n}(x,\cdot) - T^{t_n}(y,\cdot)\|_{d_n} = 0 \tag{8}$$

where $B(x,\gamma)$ is the open ball of radius γ centered at x, as measured using some metric defining the topology of S.

Then we have the following result, which is an extension of results in Hairer and Mattingly (2006) and Hairer (2008). A "reachable" point $x \in S$ means that for all open sets A containing x, $\sum_{n=1}^{\infty} T^n(y,A) > 0$ for all $y \in S$ (Meyn and Tweedie, 1993, p. 135).

Theorem 4. *Suppose that S is a Polish space and the Markov chain $\{X_n\}_{n\in\mathbb{N}}$ with transition kernel T is asymptotically strong Feller. If there is a reachable point $x \in S$ then T can have at most one stationary distribution.*

The results in Hairer (2008) require an "accessible" point, which is stronger than a reachable point. Theorem 4 is proven in Appendix A.2.

3 A Poisson Threshold Model

Our first example is a Poisson threshold model with identity link function that we have found useful in our own applications (Matteson et al., 2011). The model is defined as

$$\begin{aligned} Y_n|Y_{0:n-1} &\sim \text{Poisson}(\mu_n) \\ \mu_n &= \omega + \alpha Y_{n-1} + \beta\mu_{n-1} + (\gamma Y_{n-1} + \eta\mu_{n-1})\,\mathbf{1}_{\{Y_{n-1}\notin(L,U)\}} \end{aligned} \tag{9}$$

where the threshold boundaries satisfy $0 < L < U < \infty$. To ensure positivity of μ_n we assume $\omega, \alpha, \beta > 0$, $(\alpha+\gamma) > 0$, and $(\beta+\eta) > 0$. Additionally we take $\eta \leq 0$ and $\gamma \geq 0$, so that when Y_{n-1} is outside the range (L, U) the mean process μ_n is more adaptive, i.e. puts more weight on Y_{n-1} and less on μ_{n-1}.

We will show that a perturbed version of the model $\{Y_n\}_{n\in\mathbb{N}}$ is stationary and ergodic under the restriction $(\alpha+\beta+\gamma+\eta) < 1$. This can be proven via extension of results in Fokianos et al. (2009) for a non-threshold linear model. However, a much simpler proof is as follows. First, incorporate perturbations $Z_n \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$ as in Theorem 2:

$$\begin{aligned} Y_n^{(\sigma)}|Y_{0:n-1}^{(\sigma)} &\sim \text{Poisson}(\mu_n^{(\sigma)}) \\ \mu_n^{(\sigma)} &= \omega + \alpha Y_{n-1}^{(\sigma)} + \beta\mu_{n-1}^{(\sigma)} + \left(\gamma Y_{n-1}^{(\sigma)} + \eta\mu_{n-1}^{(\sigma)}\right)\mathbf{1}_{\{Y_{n-1}^{(\sigma)}\notin(L,U)\}} + \sigma Z_{n-1}. \end{aligned}$$

The regularity conditions for Theorem 2 hold since ψ_ν is the Poisson density and $\mu_n^{(\sigma)}$ is linear in $Z_{0:n-1}$ with bounded coefficients.

Set $X_n = \mu_n^{(\sigma)}$ and take the state space of the Markov chain $X = \{X_n\}_{n\in\mathbb{N}}$ to be $S = [\frac{\omega}{1-\beta-\eta}, \infty)$. Define $A = [\frac{\omega}{1-\beta-\eta}, \frac{\omega}{1-\beta-\eta} + M]$ for any $M > 0$, and define m to be the smallest positive integer such that $M(\beta+\eta)^{m-1} < \sigma/2$. Then

$$\begin{aligned} &\inf_{x\in A} \Pr(Y_0 = Y_1 = \ldots = Y_{m-2} = 0 | X_0 = x) > 0 \qquad \text{and} \\ &\Pr\left(\sigma(Z_0 + Z_1 + \ldots + Z_{m-2}) < \frac{\sigma}{2} - M(\beta+\eta)^{m-1}\right) > 0. \end{aligned}$$

Therefore $\inf_{x\in A} T^{m-1}(x, B) > 0$, where $B = [\frac{\omega}{1-\beta-\eta}, \frac{\omega}{1-\beta-\eta} + \frac{\sigma}{2}]$ and where T is the transition kernel of the Markov chain X. Taking $\nu = \text{Unif}(\frac{\omega}{1-\beta-\eta} + \frac{\sigma}{2}, \frac{\omega}{1-\beta-\eta} + \sigma)$ in Definition 2 then establishes A as a small set. A similar argument can be used to show φ-irreducibility and aperiodicity.

Taking the energy function $V(x) = x$,

$$\begin{aligned} E_x V(X_1) &= (\alpha+\beta)V(x) + \gamma E_x[Y_0 \mathbf{1}_{\{Y_0\notin(L,U)\}}] + \eta x P_x[Y_0 \notin (L,U)] + (\omega + \sigma/2) \\ &\leq (\alpha+\beta+\gamma)V(x) + \eta x - \eta x P_x[Y_0 \in (L,U)] + (\omega+\sigma/2). \end{aligned}$$

In particular, $E_x V(X_1)$ is bounded for $x \in A$. Also, as $x \to \infty$ we have $xP_x[Y_0 \in (L,U)] \to 0$, so for sufficiently large M, $x > M$ implies that $-\eta x P_x[Y_0 \in (L,U)] \leq 1$. Thus for $x > M$,

$$E_x V(X_1) \leq (\alpha+\beta+\gamma+\eta)V(x) + (\omega+\sigma/2+1) \leq \nu V(x)$$

for some $|\nu| < 1$ and for M large enough. So $E_x V(X_1)$ has geometric drift for $x \notin A$. Although the range of V is $[0,\infty)$ here, we can easily replace V by $\tilde{V}(x) = x + 1$ to get the range $[1,\infty)$. So the chain $\{\mu_n^{(\sigma)}\}_{n\in\mathbb{N}}$ is geometrically ergodic, and thus stationary for an appropriate initial distribution for $\mu_0^{(\sigma)}$. As shown in Section 2, this implies that the time series model $\{Y_n^{(\sigma)}\}_{n\in\mathbb{N}}$ is also stationary and ergodic. Stationarity of the original process $\{Y_n\}_{n\in\mathbb{N}}$ is addressed in Section 5.

4 Generalized Autoregressive Moving Average Models

Generalized Autoregressive Moving Average (GARMA) models are a generalization of autoregressive moving average models to exponential-family distributions, allowing direct treatment of binary and count-valued data, among others. GARMA models were stated in their most general form by Benjamin et al. (2003), based on earlier work by Zeger and Qaqish (1988) and Li (1994). Showing stationarity for GARMA models is harder than for the linear models that have been the subject of most previous studies (Bougerol and Picard, 1992; Ferland et al., 2006; Fokianos et al., 2009), since a small change in the transformed mean can correspond to a very large change on the scale of the observations, causing instability.

We write GARMA models in the following form (Benjamin et al., 2003):

$$\begin{aligned} &Y_n|Y_{0:n-1} \sim \psi_\nu(\mu_n) \\ &g(\mu_n) = \gamma + \rho[g(Y^*_{n-1}) - \gamma] + \theta[g(Y^*_{n-1}) - g(\mu_{n-1})] \end{aligned} \tag{10}$$

for some real-valued link function g, where Y^*_n is some mapping of Y_n to the domain of g, and where ψ_ν is a density function with respect to some measure on $\mathbb{R}$ (typically Lebesgue or counting measure), parameterized by ν. The second and third terms of the model (10) are the autoregressive and moving-average terms, respectively. This model is more general than the class of models developed in Benjamin et al. (2003) in the sense that we do not assume that ψ_ν is in the exponential family. However, we do assume that $E(Y_n|\mu_n) = \mu_n$, and we assume a bound on the $(2+\delta)$ moment of Y_n in terms of $|\mu_n|$, for some $\delta > 0$. We will see that our conditions are satisfied by many standard choices such as the Poisson and binomial GARMA models. In practice when applying the GARMA model covariates are often included, and multiple lags can be allowed in the autocorrelation and moving-average terms, yielding the more general model:

$$g(\mu_n) = W'_n\beta + \sum_{j=1}^{p} \rho_j[g(Y^*_{n-j}) - W'_{n-j}\beta] + \sum_{j=1}^{q} \theta_j[g(Y^*_{n-j}) - g(\mu_{n-j})]. \tag{11}$$

However, for simplicity we focus on the case $p, q = 1$; we discuss how one might extend our results to $p > 1$ and $q > 1$ at the end of Sec. 4.1. Since the covariates are time-dependent, the model (11) is in general nonstationary, and interest is in proving stationarity in the absence of covariates, i.e. where $W'_n\beta = \gamma$ as in (10).

We handle three separate cases:

Case 1: $\psi_\nu(\mu)$ is defined for any $\mu \in \mathbb{R}$. In this case the domain of g is $\mathbb{R}$ and we take $Y^*_n = Y_n$.

Case 2: $\psi_\nu(\mu)$ is defined for only $\mu \in \mathbb{R}^+$ (or μ on any one-sided open interval by analogy). In this case the domain of g is $\mathbb{R}^+$ and we take $Y^*_n = \max\{Y_n, c\}$ for some $c > 0$.

Case 3: $\psi_\nu(\mu)$ is defined for only $\mu \in (0, a)$ where $a > 0$ (or any bounded open interval by analogy). In this case the domain of g is $(0, a)$ and we take $Y^*_n = \min[\max(Y_n, c), (a - c)]$ for some $c \in (0, a/2)$.

Valid link functions g are bijective and monotonic (WLOG, increasing). Choices for Case 2 include the log link, which is the most commonly used, and the link, parameterized by $\alpha > 0$,

$$g(\mu) = \log\left(e^{\alpha\mu} - 1\right)/\alpha, \tag{12}$$

which has the property that $g(\mu) \approx \mu$ for large μ. Benjamin et al. (2003) also suggest an unmodified identity link function $g(\mu) = \mu$ for Case 2; however, this requires strong restrictions on the parameters in order to guarantee that $\mu_n \geq 0$, so we do not address this or other cases of non-surjective link functions. Examples of valid link functions for Cases 1 and 3 are the identity and logit functions, respectively.

In this section we obtain ergodicity and stationarity results for the perturbed model:

$$\begin{aligned} Y_n^{(\sigma)} | Y_{0:n-1}^{(\sigma)} &\sim \psi_\nu(\mu_n^{(\sigma)}) \\ g(\mu_n^{(\sigma)}) &= \gamma + \rho[g(Y_{n-1}^{(\sigma)*}) - \gamma] + \theta[g(Y_{n-1}^{(\sigma)*}) - g(\mu_{n-1}^{(\sigma)})] + \sigma Z_{n-1} \end{aligned} \tag{13}$$

where $Z_n \overset{\text{iid}}{\sim} N(0,1)$, for any $\sigma > 0$. Stationarity of the original model (10) is addressed in Section 5.

For the perturbed model we have the following stationarity results.

Theorem 5. *The process $\{\mu_n^{(\sigma)}\}_{n\in\mathbb{N}}$ specified by the perturbed GARMA model (13) is an ergodic Markov chain and thus stationary for an appropriate initial distribution for $\mu_0^{(\sigma)}$, under the conditions below. This implies that the perturbed GARMA model $\{Y_n^{(\sigma)}\}_{n\in\mathbb{N}}$ is stationary and ergodic when $\mu_0^{(\sigma)}$ is initialized appropriately. The conditions are:*

- $E(Y_n^{(\sigma)} | \mu_n^{(\sigma)}) = \mu_n^{(\sigma)}$
- *($2+\delta$ moment condition): There exist $\delta > 0$, $r \in [0, 1+\delta)$ and nonnegative constants d_1, d_2 such that*
$$E\left[|Y_n^{(\sigma)} - \mu_n^{(\sigma)}|^{2+\delta} \,\middle|\, \mu_n^{(\sigma)} \right] \leq d_1 |\mu_n^{(\sigma)}|^r + d_2.$$
- *g is bijective, increasing, and*

Case 1: $g : \mathbb{R} \mapsto \mathbb{R}$ is concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, and $|\rho| < 1$

Case 2: $g : \mathbb{R}^+ \mapsto \mathbb{R}$ is concave on $\mathbb{R}^+$, and $|\rho|, |\theta| < 1$

Case 3: $|\theta| < 1$; no additional conditions on $g : (0, a) \mapsto \mathbb{R}$

In fact we show the stronger condition of geometric ergodicity of the $\{\mu_n^{(\sigma)}\}_{n\in\mathbb{N}}$ process. This implies geometric ergodicity of the joint $\{(Y_n^{(\sigma)}, \mu_n^{(\sigma)})\}_{n\in\mathbb{N}}$ process, by applying Prop. 1 of Meitz and Saikkonen (2008).

The following popular models are special cases of Theorem 5:

Corollary 6. *Suppose that conditional on $\mu_n^{(\sigma)}$, $Y_n^{(\sigma)}$ is Poisson distributed with mean $\mu_n^{(\sigma)}$. Then the perturbed GARMA model is ergodic and stationary given an appropriate initial distribution for $\mu_0^{(\sigma)}$, provided that $|\rho|, |\theta| < 1$ and the link function g is bijective, increasing, and concave. This is satisfied, for instance, by the log link and the modified identity link (12). Theorem 2 applies with no further restrictions.*

Proof. If X is Poisson with mean μ then

$$E(X - \lambda)^4 = 3\lambda^2 + \lambda \leq 4\lambda^2 + 1,$$

where the inequality can be seen by considering the cases $\lambda \leq 1$ and $\lambda > 1$ separately. Thus we can take $\delta = 2$ and $r = 2$. Theorem 2 applies here, shown as follows. The Poisson density satisfies regularity condition (a). Also, $X_n = g(\mu_n^{(\sigma)})$ is linear in $Z_{0:n-1}$ and $g^{-1}(\cdot)$ is Lipschitz on any compact set (due to the concavity of g), implying that $\mu_n = g^{-1}(X_n)$ is Lipschitz in $Z_{0:n-1}$, uniformly on any compact subset of the parameter space $(\gamma, \rho, \theta) \in \mathbb{R}^3$. □

Corollary 7. *Suppose that conditional on $\mu_n^{(\sigma)}$, $Y_n^{(\sigma)}$ is binomially distributed with mean $\mu_n^{(\sigma)}$ and fixed number of trials a. Then the perturbed GARMA model is ergodic and thus stationary for an appropriate initial distribution for $\mu_0^{(\sigma)}$, provided that $|\theta| < 1$ and g is bijective and increasing (e.g. the logit link). If g^{-1} is locally Lipschitz then Theorem 2 also holds.*

The local Lipschitz condition on g^{-1} is satisfied for the logit and probit link functions, and in the case where g is differentiable holds as long as the derivative of g is nowhere zero.

Proof. The $2+\delta$ moment condition holds by taking $\delta = 0.5$ and $r = 0$:

$$E\left[|Y_n^{(\sigma)} - \mu_n^{(\sigma)}|^{2.5}\right] \leq k^{2.5}.$$

Theorem 2 applies here, by verifying the regularity conditions as for Corr. 6. Unlike the case of Corr. 6, g^{-1} is not automatically locally Lipschitz, which is why Corr. 7 explicitly makes this assumption. □

4.1 Proof of Theorem 5

Define $X_n = g(\mu_n^{(\sigma)})$; we will prove Theorem 5 by showing that the Markov chain $X = \{X_n\}_{n\in\mathbb{N}}$ with transition kernel T on state space $\mathbb{R}$ is φ-irreducible, aperiodic, and positive Harris recurrent with a geometric drift condition. Aperiodicity and φ-irreducibility are immediate since the Markov transition kernel has a (normal mixture) density that is positive on the whole real line.

Next, define the set $A = [-M, M]$ for some constant $M > 0$ to be chosen later; we will show that A is small, taking $m = 1$ and ν to be the uniform distribution on A in Definition 2. Let $x = X_0$ and write $\mu = g^{-1}(x)$. For any $y > 0$ Markov's inequality then gives

$$P_x(|Y_0^{(\sigma)} - \mu| > y) \leq \frac{E_x|Y_0^{(\sigma)} - \mu|^{2+\delta}}{y^{2+\delta}} \leq \frac{d_1|\mu|^r + d_2}{y^{2+\delta}}. \tag{14}$$

In particular, for $y = [4(d_1|\mu|^r + d_2)]^{1/(2+\delta)}$, $P_x(|Y_0^{(\sigma)} - \mu| > y) \leq 1/4$. Then for any $x \in A$,

$$\begin{aligned}
&P_x(Y_0^{(\sigma)} \in [a_1(M), a_2(M)]) > 3/4 \qquad \text{for} \\
&a_1(M) = g^{-1}(-M) - [4(d_1 \max\{|g^{-1}(-M)|, |g^{-1}(M)|\}^r + d_2)]^{1/(2+\delta)} \\
&a_2(M) = g^{-1}(M) + [4(d_1 \max\{|g^{-1}(-M)|, |g^{-1}(M)|\}^r + d_2)]^{1/(2+\delta)}.
\end{aligned}$$

Then with probability at least 3/4,

$$\begin{aligned}
&X_1 - \sigma Z_0 \geq \min\{b(a_1(M)), b(a_2(M))\} - |\theta|M \qquad \text{and} \\
&X_1 - \sigma Z_0 \leq \max\{b(a_1(M)), b(a_2(M))\} + |\theta|M \qquad \text{where} \\
&\qquad b(a) = (\rho + \theta)g(a^*) + (1 - \rho)\gamma
\end{aligned}$$

where a^* is the operator * applied to a (e.g. $a^* = \max\{a, c\}$ for Case 2). Then it is easy to see that $\exists\lambda > 0$ such that $T(x, \cdot) \geq \lambda\nu(\cdot)$ for all $x \in A$.

Next we use the small set A to prove a drift condition. Taking the energy function $V(x) = |x|$, we have the following results. First we give the drift condition for $x \in A$:

Proposition 8. ***Cases 1-3:*** *There is some constant $K(M) < \infty$ such that $E_x V(X_1) \leq K(M)$ for all $x \in A$.*

Then we give the drift condition for $x \notin A$, handling the cases $x < -M$ and $x > M$ separately:

Proposition 9. ***Cases 2-3:*** *There is some constant $K_2 < \infty$ such that $E_x V(X_1) \leq |\theta| V(x) + K_2$ for all $x < -M$.*
Case 1: *For any $\epsilon \in (0,1)$ there is some constant $K_2 < \infty$ such that for M large enough, $E_x V(X_1) \leq (|\rho| + \epsilon) V(x) + K_2$ for all $x < -M$.*

Proposition 10. ***Cases 1-2:*** *For any $\epsilon \in (0,1)$ there is some constant $K_3 < \infty$ such that for M large enough, $E_x V(X_1) \leq (|\rho| + \epsilon) V(x) + K_3$ for all $x > M$.*
Case 3: *There is some constant $K_3 < \infty$ such that $E_x V(X_1) \leq |\theta| V(x) + K_3$ for all $x > M$.*

Propositions 8-10 are proven in Appendices A.6-A.11. Propositions 9 and 10 give the overall drift condition for $x \notin A$ as follows. Consider Case 2; the other two cases are analogous. Take $\epsilon = (1 - |\rho|)/2$, define $\eta = \max\{|\theta|, |\rho| + \epsilon\} < 1$, and choose M large enough to satisfy Prop. 10. Then for any $x \notin A$ we have

$$\begin{aligned} E_x V(X_1) &\leq \eta V(x) + \max\{K_2, K_3\} \\ &\leq \frac{\eta + 1}{2} V(x) \end{aligned}$$

for M large enough, establishing geometric ergodicity (although the range of V is $[0, \infty)$, we can easily replace V with $\tilde{V}(x) = |x| + 1$ to get the range $[1, \infty)$).

These results have the following intuition for Case 2: Prop. 9 shows that for very negative X_{n-1}, $|\theta|$ controls the rate of drift, while Prop. 10 shows that for large positive X_{n-1}, $|\rho|$ controls the rate of drift. The former result is due to the fact that for very negative values of X_{n-1} the autoregressive term in (13) is a constant, $\rho(g(c) - \gamma)$, so the moving-average term dominates. The latter result is due to the fact that for large positive X_{n-1}, the distribution of $Y_{n-1}^{(\sigma)}$ concentrates around $\mu_{n-1}^{(\sigma)}$, so that the moving-average term $\theta[g(Y_{n-1}^{(\sigma)*}) - g(\mu_{n-1}^{(\sigma)})]$ in (13) is negligible and the autoregressive term dominates.

For a perturbed version of the GARMA model with multiple lags (11), it may be possible to show geometric ergodicity of the multivariate Markov chain with state vector $\mu_{(n-\max\{p,q\}+1):n}^{(\sigma)}$. Again this could be done by finding a small set and energy function such that a drift condition holds, subject to appropriate restrictions on the parameters $(\rho_1, \ldots, \rho_p)$ and $(\theta_1, \ldots, \theta_q)$.

5 Stationarity of the Original Model

In this section we will show existence and uniqueness of the stationary distribution for the Poisson threshold model and the class of GARMA models. These results potentially form the foundation for broadly showing consistency and asymptotic normality of maximum likelihood estimators in these models.

5.1 The Poisson Threshold Model

We will illustrate the use of Feller properties to show that a discrete-valued time series model has a unique stationary distribution. For the Poisson threshold model (9) we first show existence of a stationary distribution.

Lemma 11. *The Markov chain $\{\mu_n\}_{n \in \mathbb{N}}$ defined by (9) has a stationary distribution, under the restriction $(\alpha + \beta + \gamma + \eta) < 1$.*

Proof. We use Theorem 3. The space $S = [\frac{w}{1-\beta-\eta}, \infty)$ is a locally compact complete separable metric space with Borel σ-field. Let $Y_0(x)$ and $\mu_1(x)$ denote the random variables Y_0 and μ_1 conditioned on $\mu_0 = x$. Since $Y_0(x) = \text{Pois}(x)$ we have that $Y_0(x)$ converges in distribution to $Y_0(y)$ as $x \to y$ for any $y \in S$. Therefore $\mu_1(x)$ converges in distribution to $\mu_1(y)$ as $x \to y$, proving that the chain $\{\mu_n\}_{n\in\mathbb{N}}$ is weak Feller. The set A defined in Section 3 is compact, and a drift condition for this set is shown in that Section (the proof of the drift condition is valid in the case $\sigma = 0$). By Theorem 3 the chain $\{\mu_n\}_{n\in\mathbb{N}}$ has a stationary distribution. □

Next we show uniqueness of the stationary distribution.

Lemma 12. *The chain $\{\mu_n\}_{n\in\mathbb{N}}$ defined by (9) has at most one stationary distribution, provided that $\beta < 1$.*

Proof. The space S is a Polish space. The point $x = \frac{w}{1-\beta-\eta}$ is reachable, shown as follows. For any initial state $\mu_0 = y$ and any m the probability that $Y_n = 0$ for all $n = 0, \ldots, m$ is strictly positive. So for any open set A containing x we can choose m large enough that $\mu_m \in A$ with positive probability; therefore x is reachable.

The process $\{\mu_n\}_{n\in\mathbb{N}}$ is asymptotically strong Feller; the proof is given in Appendix A.5, under the condition $\beta < 1$. Theorem 4 then implies that the process $\{\mu_n\}$ has at most one stationary distribution. □

Putting together Lemmas 11 and 12 we find that:

Corollary 13. *The mean process $\{\mu_n\}_{n\in\mathbb{N}}$ of the Poisson threshold model defined by (9), under the restrictions $(\alpha + \beta + \gamma + \eta) < 1$ and $\beta < 1$, has a unique stationary distribution π. When μ_0 is initialized according to π the Poisson threshold model $\{Y_n\}_{n\in\mathbb{N}}$ is strictly stationary.*

5.2 The GARMA Model

First we show existence of a stationary distribution for the GARMA model (10) by using the weak Feller property. Let $Y_0(x)$ denote the random variable Y_0 conditioned on $\mu_0 = x$.

Theorem 14. *The process $\{\mu_n\}_{n\in\mathbb{N}}$ specified by the GARMA model (10) has a stationary distribution, and thus is stationary for an appropriate initial distribution for μ_0, under the conditions below. This implies that the GARMA model $\{Y_n\}_{n\in\mathbb{N}}$ is stationary when μ_0 is initialized appropriately. The conditions are:*

- *$Y_0(x) \Rightarrow Y_0(y)$ as $x \to y$*
- *$E(Y_n|\mu_n) = \mu_n$*
- *($2 + \delta$ moment condition): There exist $\delta > 0$, $r \in [0, 1 + \delta)$ and nonnegative constants d_1, d_2 such that*
$$E\left[|Y_n - \mu_n|^{2+\delta} \,\middle|\, \mu_n\right] \le d_1|\mu_n|^r + d_2.$$
- *g is bijective, increasing, and*

Case 1: $g : \mathbb{R} \mapsto \mathbb{R}$ is concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, and $|\rho| < 1$

Case 2: $g : \mathbb{R}^+ \mapsto \mathbb{R}$ is concave on $\mathbb{R}^+$, and $|\rho|, |\theta| < 1$

Case 3: $|\theta| < 1$; no additional conditions on $g : (0, a) \mapsto \mathbb{R}$.

Proof. We apply Theorem 3 to the chain $\{g(\mu_n)\}_{n\in\mathbb{N}}$ to show that it has a stationary distribution; this implies the same result for the chain $\{\mu_n\}_{n\in\mathbb{N}}$. The state space $S = \mathbb{R}$ of $\{g(\mu_n)\}_{n\in\mathbb{N}}$ is a locally compact complete separable metric space with Borel σ-field. A drift condition for $\{g(\mu_n)\}_{n\in\mathbb{N}}$ is given in the proof of Theorem 5, for the compact set $A = [-M, M]$ (the proof of that drift condition holds when $\sigma = 0$). All that remains is to show that the chain $\{g(\mu_n)\}_{n\in\mathbb{N}}$ is weak Feller. Let $X_n = g(\mu_n)$. For $X_0 = x$ we have that

$$X_1(x) = \gamma + \rho(g(Y_0^*(g^{-1}(x))) - \gamma) + \theta(g(Y_0^*(g^{-1}(x))) - x).$$

Since g^{-1} is continuous, $Y_0(g^{-1}(x)) \Rightarrow Y_0(g^{-1}(y))$ as $x \to y$. Since the * operation that maps Y_0 to the domain of g is continuous, it follows that $Y_0^*(g^{-1}(x)) \Rightarrow Y_0^*(g^{-1}(y))$ as $x \to y$. Since g is continuous, we have that $g(Y_0^*(g^{-1}(x))) \Rightarrow g(Y_0^*(g^{-1}(y)))$. So $X_1(x) \Rightarrow X_1(y)$ as $x \to y$, showing the weak Feller property. □

Next we show uniqueness of the stationary distribution, using the asymptotic strong Feller property. We will assume that the distribution $\pi_z(\cdot)$ of $g(Y_n^*)$ conditional on $g(\mu_n) = z$ varies smoothly and not too quickly as a function of z. By this we mean that $\pi_z(\cdot)$ has the Lipschitz property

$$\sup_{w,z\in\mathbb{R}:w\neq z} \frac{\|\pi_w(\cdot) - \pi_z(\cdot)\|_{TV}}{|w - z|} < B < \infty. \tag{15}$$

Theorem 15. *Suppose that the conditions of Thm. 14 and the Lipschitz condition (15) hold, and that there is some $x \in \mathbb{R}$ that is in the support of Y_0 for all values of μ_0. Then there is a unique stationary distribution for $\{\mu_n\}_{n\in\mathbb{N}}$.*

This result is proven in Appendix A.3.

The following two results give two classes of examples where Theorem 15 may be applied. The proofs of these results may be found in Appendix A.4.

Proposition 16. *Suppose that conditional on μ_n, Y_n is Poisson(μ_n), the link function $g : \mathbb{R}^+ \to \mathbb{R}$ is bijective, concave and increasing, g^{-1} is Lipschitz, $|\rho|, |\theta| < 1$ and $c \in (0, 1)$. Then the process $\{\mu_n\}_{n\in\mathbb{N}}$ defined in (10) has a unique stationary distribution π. Hence, when μ_0 is initialized according to π, the process $\{Y_n\}_{n\in\mathbb{N}}$ is strictly stationary.*

The condition that g^{-1} be Lipschitz is equivalent to requiring that for $y > x$, $g(y) - g(x) \geq \zeta(y - x)$ for some $\zeta > 0$, and this condition in turn is equivalent to requiring that g' be bounded away from zero when g is differentiable. The link function (12) satisfies this condition, while the log link does not.

We suspect that the Lipschitz condition in Theorem 15 can be weakened to a local Lipschitz condition, based on the fact that local Lipschitz is equivalent to Lipschitz on a compact space, and the fact that although the state space of $g(\mu_n)$ is not compact, we have a drift condition for the process $\{g(\mu_n)\}_{n\in\mathbb{N}}$ which (informally) ensures that the chain stays in a limited part of the space. With the weaker local Lipschitz condition, Proposition 16 could be extended to link functions like the log link.

Proposition 17. *Suppose that conditional on μ_n, Y_n is binomial with fixed number of trials a and mean μ_n, the link function $g : (0, a) \to \mathbb{R}$ is bijective and increasing, g^{-1} is Lipschitz, $|\theta| < 1$ and $c \in (0, 1)$. Then the process $\{\mu_n\}_{n\in\mathbb{N}}$ defined in (10) has a unique stationary distribution π. Hence, when μ_0 is initialized according to π, the process $\{Y_n\}_{n\in\mathbb{N}}$ is strictly stationary.*

A Appendix: Proofs

A.1 Proof of Theorem 2

Fixing $y_{0:n}$ and letting $Z = Z_{0:n-1}$ be the perturbations,

$$\sup_{\eta\in K} |\mathcal{L}_\sigma(\eta) - \mathcal{L}(\eta)| = \sup_{\eta\in K} \left| E \prod_{k=0}^{n} \psi_\nu(y_k; \mu_k(\sigma Z)) - \prod_{k=0}^{n} \psi_\nu(y_k; \mu_k(0)) \right|$$

where the expectation is taken over Z, the data being fixed. Then we have

$$\begin{aligned}
\sup_{\eta\in K} |\mathcal{L}_\sigma(\eta) - \mathcal{L}(\eta)| \leq & \sup_{\eta\in K} E \left| \prod_{k=0}^{n} \psi_\nu(y_k; \mu_k(\sigma Z)) - \prod_{k=0}^{n} \psi_\nu(y_k; \mu_k(0)) \right| \\
\leq & E \sup_{\eta\in K} \left| \prod_{k=0}^{n} \psi_\nu(y_k; \mu_k(\sigma Z)) - \prod_{k=0}^{n} \psi_\nu(y_k; \mu_k(0)) \right| \\
= & E \sup_{\eta\in K} \left| \prod_{k=0}^{n} \beta_k(\sigma Z) - \prod_{k=0}^{n} \beta_k(0) \right| \qquad (16)
\end{aligned}$$

where $\beta_k(\cdot) = \psi_\nu(y_k; \mu_k(\cdot))$. We will show that the supremum inside the expectation in (16) converges to 0 almost surely (in Z) as $\sigma \to 0$; then bounded convergence implies that the expectation (16) itself converges to 0 as $\sigma \to 0$, proving Thm. 2.

By assumption the function $\psi_\nu(y; \mu)$ is Lipschitz continuous in μ, and $\mu_k(\cdot)$ is Lipschitz continuous in some bounded neighborhood C of 0, uniformly in $\eta \in K$. In other words, there exists a finite constant L_k such that, for any $z, z' \in C$,

$$\sup_{\eta\in K} |\mu_k(z) - \mu_k(z')| \leq L_k \|z - z'\|$$

for each $k = 0, 1, \ldots, n$. Thus, the composition $\beta_k(\cdot) = \psi_\nu(y_k, \mu_k(\cdot))$ is Lipschitz continuous on C, uniformly in $\eta \in K$, for each $k = 0, 1, \ldots, n$.

Finally, we apply the usual telescoping-sum argument to conclude that the function $\prod_{k=0}^{n} \beta_k(\cdot)$ is Lipschitz in $z \in C$, uniformly in $\eta \in K$. For any $z, z' \in C$,

$$\begin{aligned}
\left| \prod_{k=0}^{n} \beta_k(z) - \prod_{k=0}^{n} \beta_k(z') \right| &= \left| \sum_{k=0}^{n} \left(\prod_{i=0}^{n-k} \beta_i(z) \prod_{j=n-k+1}^{n} \beta_j(z') - \prod_{i=0}^{n-k-1} \beta_i(z) \prod_{j=n-k}^{n} \beta_j(z') \right) \right| \\
&= \left| \sum_{k=0}^{n} (\beta_{n-k}(z) - \beta_{n-k}(z')) \prod_{i=0}^{n-k-1} \beta_i(z) \prod_{j=n-k+1}^{n} \beta_j(z') \right| \\
&\leq \sum_{k=0}^{n} \left[\prod_{j\neq n-k} \sup_{\mu} \psi_\nu(y_j; \mu) \right] |\beta_{n-k}(z) - \beta_{n-k}(z')|.
\end{aligned}$$

By regularity condition (a), $\left[\prod_{j\neq n-k} \sup_\mu \psi_\nu(y_j; \mu) \right]$ is bounded uniformly in $\eta \in K$ for each k. The fact that $\beta_k(\cdot)$ is Lipschitz uniformly in $\eta \in K$ for each $k = 0, 1, \ldots, n$ then ensures that $\prod_{k=0}^{n} \beta_k(\cdot)$ is Lipschitz on C, uniformly in $\eta \in K$ as desired.

A.2 Proof of Theorem 4

We will show that if a point x is reachable, then x is in the support of any invariant distribution. Combined with Corollary 3.17 of Hairer and Mattingly (2006) this gives the desired result.

Let x be reachable and let π be an invariant probability measure. We will show that x is in the support of π by showing that $\pi(A) > 0$ for all open sets A containing x (see Lemma 3.7 of Hairer and Mattingly 2006). To begin, let A be an arbitrary open set containing x. Let B_n be the set of initial states $y \in S$ that have positive probability of hitting A on the nth step, i.e., $B_n = \{y : T^n(y, A) > 0\}$.

Since x is reachable, the countable union of $\{B_n : n \geq 1\}$ contains S. Since $\pi(S) = 1$, it follows that the π measure of at least one B_n is strictly positive. Fix $n \geq 1$ such that this is the case. Then

$$\pi(A) = \int_S \pi(dy)T^n(y, A) \geq \int_{B_n} \pi(dy)T^n(y, A) > 0.$$

The fact that this quantity is strictly positive follows using a standard argument as follows. First, we can write B_n as the countable union of the increasing sets C_k, $k \geq 1$, where

$$C_k = \{y : T^n(y, A) \geq 1/k\}.$$

So then $\pi(B_n) = \lim_{k\to\infty} \pi(C_k)$. Fix $k > 0$ such that $\pi(C_k) > 0$. Then

$$\int_{B_n} \pi(dy)T^n(y, A) \geq \int_{C_k} \pi(dy)T^n(y, A) \geq \int_{C_k} \pi(dy)\frac{1}{k} = \frac{\pi(C_k)}{k} > 0.$$

A.3 Proof of Theorem 15

Let $Z_n = g(\mu_n)$ for all $n \geq 0$. We will show that the Markov chain $\{Z_n\}_{n\in\mathbb{N}}$ is asymptotically strong Feller, and that there is a reachable point for $\{Z_n\}_{n\in\mathbb{N}}$. Combined with Theorem 4 and the fact that $\mathbb{R}$ is Polish this shows that the chain $\{Z_n\}$ can have at most one stationary distribution. Since g is bijective, this implies that the Markov chain $\{\mu_n\}_{n\in\mathbb{N}}$ can have at most one stationary distribution. Combined with Theorem 14 this gives the desired result.

First we show the existence of a reachable point for $\{Z_n\}_{n\in\mathbb{N}}$. Since there is some point $x \in \mathbb{R}$ that is in the support of Y_n for all values of μ_n, the point $g(x^*)$ is in the support of $g(Y_n^*)$ for all values of μ_n (since the transformations * and g are continuous and monotonic). So for every open set $B \ni g(x^*)$ we have $\Pr(g(Y_n^*) \in B \mid \mu_n) > 0$ for all μ_n. Furthermore, $\Pr(g(Y_j^*) \in B \; \forall j = 1, \ldots, n \mid \mu_0) > 0$.

We can rewrite the definition of $g(\mu_n)$ given in (10) as

$$g(\mu_n) = \gamma(1-\rho)\sum_{j=0}^{n-1}(-\theta)^j + (\rho+\theta)\sum_{j=0}^{n-1}(-\theta)^j g(Y_{n-1-j}^*) + (-\theta)^n g(\mu_0).$$

We will show that $z = [\gamma(1-\rho) + (\rho+\theta)g(x^*)]/(1+\theta)$ is a reachable point, using the fact that $\sum_{j=0}^{n-1}(-\theta)^j \overset{n\to\infty}{\longrightarrow} 1/(1+\theta)$. Take any open set $A \ni z$, and any initial value μ_0; we will show that $\exists n$ such that $\Pr(g(\mu_n) \in A \mid \mu_0) > 0$. Letting $B(z, \epsilon)$ indicate the open ball of radius $\epsilon > 0$ centered at z, there is some ϵ such that $B(z, \epsilon) \subset A$. Then for some $\delta > 0$ we have that for all $w \in B(g(x^*), \delta)$,

$$[\gamma(1-\rho) + (\rho+\theta)w]/(1+\theta) \in B(z, \epsilon/2).$$

Choose n large enough that for all $w \in B(g(x^*), \delta)$ we have

$$[\gamma(1-\rho) + (\rho+\theta)w] \sum_{j=0}^{n-1} (-\theta)^j + (-\theta)^n g(\mu_0) \in B(z, \epsilon) \subset A.$$

Since $\Pr\big[g(Y_j^*) \in B(g(x^*), \delta)\ \forall j = 1, \ldots, n-1 \mid \mu_0\big] > 0$, we have $\Pr(g(\mu_n) \in A \mid \mu_0) > 0$ as desired.

To show that $\{Z_n\}_{n\in\mathbb{N}}$ is asymptotically strong Feller we will use the sequence of metrics d_n defined by

$$d_n(x,y) = \begin{cases} n|x-y| & |x-y| < 1/n \\ 1 & \text{else.} \end{cases} \tag{17}$$

By Example 3.2 (1) in Hairer and Mattingly (2006) this is a totally separating system of metrics. We will also define $t_n = n$.

An interesting property of the distance metric (7) is that if we take $d(x,y) = \mathbf{1}_{\{x\neq y\}}$ then we get the total variation distance between the probability measures μ_1, μ_2. This is because in this case taking the supremum over $\{\phi : \mathrm{Lip}_d \phi = 1\}$ is equivalent to taking the supremum over $\{\phi : \phi(x) \in [0,1]\ \forall x \in \mathbb{R}\}$. An analogous result is true for our choice of distance d_n, that when taking the supremum over $\{\phi : \mathrm{Lip}_{d_n} \phi = 1\}$ it is sufficient to consider ϕ such that $\phi(x) \in [0,1]\ \forall x \in \mathbb{R}$ and $\mathrm{Lip}_{d_n}\phi = 1$.

Let $Y_n(z)$ and $Z_n(z)$ indicate the random variables Y_n and Z_n conditioned on $Z_0 = z$. By (15) we have $\|\pi_z(\cdot) - \pi_w(\cdot)\|_{TV} < B|z-w|$. Using Proposition 3(g) of Roberts and Rosenthal (2004) we can construct the random variables $g(Y_0^*(z))$ and $g(Y_0^*(w))$ in such a way that they have the correct marginal distributions π_z and π_w, and that $\Pr(g(Y_0^*(w)) = g(Y_0^*(z))) \geq 1 - \|\pi_w(\cdot) - \pi_z(\cdot)\|_{TV} > 1 - B|z-w|$.

If $g(Y_0^*(w)) = g(Y_0^*(z))$ then $|Z_1(w) - Z_1(z)| = |\theta||z-w|$, and so $\|\pi_{Z_1(z)}(\cdot) - \pi_{Z_1(w)}(\cdot)\|_{TV} < |\theta||z-w|B$. Then we can construct $g(Y_1^*(z))$ and $g(Y_1^*(w))$ so that they have the correct marginal distributions, and that $\Pr(g(Y_1^*(z)) = g(Y_1^*(w)) \mid g(Y_0^*(w)) = g(Y_0^*(z))) \geq 1 - \|\pi_{Z_1(z)}(\cdot) - \pi_{Z_1(w)}(\cdot)\|_{TV} > 1 - |\theta||z-w|B$. If $g(Y_1^*(z)) = g(Y_1^*(w))$ then we can continue to "couple" the chains in the above way. Notice that the probability that the chains couple for all times $0, 1, \ldots$ is at least $1 - B|z-w| \sum_{n=0}^{\infty} |\theta|^n = 1 - \frac{|z-w|B}{1-|\theta|}$.

Consider the distance $\|T^n(z,\cdot) - T^n(w,\cdot)\|_{d_n}$; we will bound this by conditioning on whether or not the chains couple for all time. If they couple for all time, then $|Z_n(z) - Z_n(w)| = |\theta|^n|z-w|$. Due to this fact and the fact that it is sufficient to consider ϕ such that $\phi(x) \in [0,1]$

for all $x \in \mathbb{R}$,

$$
\begin{aligned}
&\|T^{t_n}(z,\cdot) - T^{t_n}(w,\cdot)\|_{d_n} \\
&= \|T^{n}(z,\cdot) - T^{n}(w,\cdot)\|_{d_n} \\
&= \sup_{\mathrm{Lip}_{d_n}\phi=1} \left(\int \phi(x) T^n(z,dx) - \int \phi(x) T^n(w,dx) \right) \\
&= \sup_{\mathrm{Lip}_{d_n}\phi=1} \left(E[\phi(Z_n(z))] - E[\phi(Z_n(w))] \right) \\
&= \sup_{\mathrm{Lip}_{d_n}\phi=1} E[\phi(Z_n(z)) - \phi(Z_n(w))] \\
&\le \sup_{\mathrm{Lip}_{d_n}\phi=1} E\big[\phi(Z_n(z)) - \phi(Z_n(w)) \;\big|\; g(Y^*_{0:n}(z)) = g(Y^*_{0:n}(w))\big] + \\
&\qquad \sup_{\mathrm{Lip}_{d_n}\phi=1} E\big[\phi(Z_n(z)) - \phi(Z_n(w)) \;\big|\; g(Y^*_{0:n}(z)) \neq g(Y^*_{0:n}(w))\big] \times \Pr\big[g(Y^*_{0:n}(z)) \neq g(Y^*_{0:n}(w))\big] \\
&\le n|\theta|^n |z-w| + \Pr\big[g(Y^*_{0:n}(z) \neq g(Y^*_{0:n}(w)\big] \\
&\le n|\theta|^n |z-w| + \frac{|z-w|B}{1-|\theta|}.
\end{aligned}
$$

So

$$
\limsup_{n\to\infty} \sup_{y \in B(x,\gamma)} \|T^{t_n}(x,\cdot) - T^{t_n}(y,\cdot)\|_{d_n} \le \frac{\gamma B}{1-|\theta|}
$$

which converges to 0 as $\gamma \to 0$. Therefore the process $\{Z_n\}_{n\in\mathbb{N}}$ is asymptotically strong Feller.

A.4 Proof of Propositions 16 and 17

In view of Corollary 6 it suffices to verify the two conditions stated in Theorem 15 to prove Proposition 16.

Zero is in the support of Y_0 for all values of μ_0. To establish the Lipschitz condition (15), we use coupling theory as follows. Let $Y_n(z)$ denote Y_n conditional on $g(\mu_n) = z$. Suppose that z and w are two values of $g(\mu_n)$. Then the total variation distance between $g(Y_n^*(z))$ and $g(Y_n^*(w))$ is

$$
\begin{aligned}
d_{TV}(g(Y_n^*(z)), g(Y_n^*(w))) &= \sup_A |P(g(Y_n^*(z)) \in A) - P(g(Y_n^*(w)) \in A)| \\
&= \sup_A |P(Y_n^*(z) \in A) - P(Y_n^*(w) \in A)| \\
&= \sup_A |P(Y_n(z) \in A) - P(Y_n(w) \in A)| \\
&= d_{TV}(Y_n(z), Y_n(w)),
\end{aligned}
$$

since g is invertible, and we can recover Y_n from Y_n^* since $c \in (0,1)$.

The coupling inequality, e.g., Thorisson (1995), ensures that

$$
d_{TV}(X,Y) \le P(X' \neq Y')
$$

for any random variables X' and Y' such that $X' \overset{\mathcal{D}}{=} X$ and $Y' \overset{\mathcal{D}}{=} Y$, where $\overset{\mathcal{D}}{=}$ means "has the same distribution as." The key point is that the joint distribution of X' and Y' is arbitrary. We choose X' and Y' in such a way that we can bound $P(X' \neq Y')$ and therefore obtain a

bound on the total-variation distance between X and Y. When this bound is Lipschitz, we then have the desired property.

So, suppose $z > w$. Let $Y_n(w)$ be Poisson distributed with mean $g^{-1}(w)$. Let ξ be a Poisson random variable, independent of $Y_n(w)$, with mean $g^{-1}(z)-g^{-1}(w)$, and set $Y_n(z) = Y_n(w)+\xi$. Then

$$\begin{aligned} P(Y_n(z) \neq Y_n(w)) &= P(\xi > 0) \\ &= 1 - \exp(-[g^{-1}(z) - g^{-1}(w)]). \end{aligned} \tag{18}$$

Let ζ be the Lipschitz constant for g^{-1}. Then (18) is bounded above by

$$1 - \exp(-\zeta(z - w))$$

which is Lipschitz, with Lipschitz constant ζ, and this completes the proof of Proposition 16.

The proof of Proposition 17 follows exactly the same lines as Proposition 16 except for the coupling used. To this end, suppose that $z > w$ and let $Y_n(z)$ be binomially distributed with parameters a (number of trials) and $g^{-1}(z)/a$ (probability of success). Let $Y_n(w)$ be conditionally binomially distributed with parameters $Y_n(z)$ and $g^{-1}(w)/g^{-1}(z)$, conditional on $Y_n(w)$. Then $Y_n(w)$ is (marginally) binomially distributed with mean $g^{-1}(w)$, and

$$P(Y_n(z) \neq Y_n(w)) = E\left(1 - \left(\frac{g^{-1}(w)}{g^{-1}(z)}\right)^{Y_n(z)}\right).$$

The moment generating function of a binomial random variable $X \sim \text{Bin}(a, p)$ with a trials and probability p of success is $E(e^{tX}) = (pe^t + 1 - p)^a$. Taking $e^t = g^{-1}(w)/g^{-1}(z)$ gives

$$\begin{aligned} P(Y_n(z) \neq Y_n(w)) &= 1 - \left(1 - \frac{g^{-1}(z) - g^{-1}(w)}{a}\right)^a \\ &\leq 1 - \left(1 - \min\left\{\frac{\zeta(z-w)}{a}, 1\right\}\right)^a \\ &= 1 - \left(\max\left\{1 - \frac{\zeta(z-w)}{a}, 0\right\}\right)^a. \end{aligned} \tag{19}$$

Now, the function $(1 - \zeta(z - w)/a)^a$ is Lipschitz for $z \in [w, w + a/\zeta]$ as can be seen since the absolute value of its derivative is bounded (by ζ), and this implies that (19) is Lipschitz. This completes the proof of Proposition 17.

A.5 Proof that Model (9) is Asymptotically Strong Feller

The proof is nearly identical to that for the GARMA model given in Appendix A.3. However, it requires that $1 > \max\{\beta + \eta, \beta\} = \beta$. The necessary Lipschitz property referred to in that proof holds for the Poisson threshold model (9) since this model uses the identity link function.

To give more detail, let $Z_n = \mu_n$ and let $\pi_z(\cdot)$ be the distribution of Y_n conditional on $Z_n = z$, i.e. $\pi_z = \text{Pois}(z)$. The proof of Prop. 16 then implies that the Lipschitz condition (15) holds. As in Appendix A.3, use the system of metrics d_n defined in (17) and define $t_n = n$. Let $Y_n(z)$ and $\mu_n(z)$ indicate the random variables Y_n and μ_n conditioned on $\mu_0 = z$. We have $\|\pi_z(\cdot)-\pi_w(\cdot)\|_{TV} < B|z-w|$. So we can construct $Y_0(z)$ and $Y_0(w)$ in such a way that they have

the correct marginal distributions π_z and π_w, and $\Pr(Y_0(z) = Y_0(w)) \geq 1 - \|\pi_z(\cdot) - \pi_w(\cdot)\|_{TV} > 1 - B|z - w|$. If $Y_0(z) = Y_0(w)$ then

$$\begin{aligned} |\mu_1(w) - \mu_1(z)| &= \begin{cases} \beta|w - z| & Y_0 \in (L, U) \\ (\beta + \eta)|w - z| & \text{else} \end{cases} \\ &\leq \beta|w - z|. \end{aligned}$$

This implies that $\|\pi_{\mu_1(w)}(\cdot) - \pi_{\mu_1(z)}(\cdot)\|_{TV} < B\beta|w - z|$. If $\beta < 1$, the probability that the chains "couple" in this way for all time is at least $1 - \frac{B|w-z|}{1-\beta}$. The rest of the argument from Sec A.3 holds unchanged.

A.6 Proof of Proposition 8, Case 1

For readability we make the dependence of $Y_0^{(\sigma)}$ on σ implicit. Recall that $\mu = g^{-1}(x)$, and assume WLOG that $g(0) = 0$, since replacing $g(y)$ with $h(y) = g(y) - g(0)$ simply changes the value of γ. Due to the fact that g is concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, there are constants $a_0, a_1 \geq 0$ such that $|g(y)| \leq a_0 + a_1|y|$ for all y. Using these facts, equation (13), and the triangle inequality, we can bound $E_x V(X_1)$ as follows, where d_i denote bounded (in μ) constants for each $i \geq 3$:

$$\begin{aligned} E_x V(X_1) &= E_x|(1 - \rho)\gamma + \rho g(Y_0) + \theta(g(Y_0) - x) + \sigma Z_0| \\ &\leq (1 - \rho)|\gamma| + \sqrt{2\sigma^2/\pi} + |\rho| E_x|g(Y_0)| + |\theta| E_x|g(Y_0) - x| \\ &\leq d_3 + (|\rho| + |\theta|)a_1 E_x|Y_0| + |\theta||x|. \end{aligned} \tag{20}$$

By the triangle and Jensen's inequalities,

$$\begin{aligned} E_x|Y_0| &= E_x|\mu + Y_0 - \mu| \\ &\leq |\mu| + E_x|Y_0 - \mu| \\ &\leq |\mu| + \left[E_x|Y_0 - \mu|^{2+\delta}\right]^{1/(2+\delta)} \\ &\leq |\mu| + (d_1|\mu|^r + d_2)^{1/(2+\delta)}. \end{aligned} \tag{21}$$

So $\sup_{x \in [-M,M]} E_x V(X_1) < \infty$, proving Prop. 8.

A.7 Proof of Propositions 9 and 10, Case 1

We will prove Prop. 10 for Case 1; Prop. 9 for Case 1 then holds by symmetry. We will show that for large x, the autoregressive part of the GARMA model dominates and the moving-average portion of the model is negligible. In the bound (20), the autoregressive part of the model is captured by $|\rho| E_x|g(Y_0)|$, while the moving-average part corresponds to the term $|\theta| E_x|g(Y_0) - x|$. Since $g(0) = 0$ and g is monotonic increasing, for all x large enough

$$\begin{aligned} E_x|g(Y_0)| &= E_x[g(Y_0)\mathbf{1}_{Y_0 > 0}] - E_x[g(Y_0)\mathbf{1}_{Y_0 < 0}] \\ &= E_x g(Y_0\mathbf{1}_{Y_0 > 0}) - E_x g(Y_0\mathbf{1}_{Y_0 < 0}) \\ &\leq g(E_x[Y_0\mathbf{1}_{Y_0 > 0}]) - g(E_x[Y_0\mathbf{1}_{Y_0 < 0}]) \\ &= g(E_x Y_0 - E_x[Y_0\mathbf{1}_{Y_0 < 0}]) - g(E_x[Y_0\mathbf{1}_{Y_0 < 0}]) \end{aligned} \tag{22}$$

by Jensen's inequality. Now, $\mu = g^{-1}(x) > 0$ for $x > 0$, so using (14)

$$\begin{aligned} -E_x[Y_0 \mathbf{1}_{Y_0<0}] &= \int_0^\infty P_x(Y_0 < -u)\, du \\ &\leq \int_0^\infty P_x(|Y_0 - \mu| > u + \mu)\, du \\ &\leq \int_0^\infty \frac{d_1\mu^r + d_2}{(u+\mu)^{2+\delta}}\, du \\ &= \frac{d_1\mu^r + d_2}{(1+\delta)\mu^{1+\delta}} \ \to 0 \end{aligned} \tag{23}$$

as $x \to \infty$. Thus, from (22), for any given $\epsilon > 0$, there exists $M > 0$ so that for $x > M$,

$$E_x|g(Y_0)| \leq g(E_xY_0 + \epsilon) + \epsilon \leq g(E_xY_0) + g(\epsilon) + \epsilon = x + d_4 \tag{24}$$

where the second inequality is due to concavity of g on $\mathbb{R}^+$.

Next we show that the term $E_x|g(Y_0) - x|$ in (20) is "small" relative to the linear (in x) term:

Proposition 18. *There is some constant d_{13} such that*

$$E_x|g(Y_0) - x| \leq d_{13}x^{r/(2+\delta)}$$

for all x large enough.

Prop. 18 is proven in Appendix A.11. Combining it with (20) and (24), we have that for all x large enough,

$$\begin{aligned} E_xV(X_1) &\leq d_{14} + |\rho|x + |\theta|d_{13}x^{r/(2+\delta)} \\ &\leq d_{14} + (|\rho| + \epsilon)x \end{aligned}$$

proving Prop. 10. □

A.8 Proof of Proposition 8 and Proposition 9, Case 2

Assume WLOG that $g(c) = 0$, since replacing $g(y)$ with $h(y) = g(y) - g(c)$ simply changes the value of γ. Since $g(c) = 0$, $g(Y_0^*) \geq 0$ is nonnegative for any Y_0^*. Also, due to the concavity of g, there is some $a_1 > 0$ such that $g(y) \leq a_1 y$ for all $y \in \mathbb{R}^+$. Using these facts, equation (13), and the triangle inequality, we can bound $E_xV(X_1)$ as follows:

$$\begin{aligned} E_xV(X_1) &= E_x|(1-\rho)\gamma + \rho g(Y_0^*) + \theta(g(Y_0^*) - x) + \sigma Z_0| \\ &\leq (1-\rho)|\gamma| + \sqrt{2\sigma^2/\pi} + |\rho|E_x[g(Y_0^*)] + |\theta|E_x|g(Y_0^*) - x| \\ &= d_{15} + |\rho|P_x(Y_0 < c)g(c) + |\rho|E_x[g(Y_0)\mathbf{1}_{Y_0\geq c}] + \\ &\quad |\theta|P_x(Y_0 < c)|g(c) - x| + |\theta|E_x[|g(Y_0) - x|\mathbf{1}_{Y_0\geq c}] \\ &\leq d_{15} + (|\rho| + |\theta|)E_x[g(Y_0)\mathbf{1}_{Y_0\geq c}] + \\ &\quad |\theta|P_x(Y_0 < c)|g(c) - x| + |\theta|P_x(Y_0 \geq c)|x| \\ &\leq d_{15} + (|\rho| + |\theta|)a_1E_x[Y_0\mathbf{1}_{Y_0\geq c}] + |\theta||x| \end{aligned} \tag{25}$$

In the same way that we obtained (21) for Case 1, we have the following bound for Case 2:

$$\begin{aligned} E_x[Y_0 \mathbf{1}_{Y_0 \geq c}] &\leq E_x|Y_0| \leq \mu + (d_1 \mu^r + d_2)^{1/(2+\delta)} \\ &\leq d_{16} + d_{17} \mu^{r/(2+\delta)} \end{aligned}$$

where $\mu = g^{-1}(x)$, implying that

$$E_x V(X_1) \leq d_{18} + d_{19}\mu + |\theta|\,|x|.$$

This is sufficient to get a uniform bound on $E_x V(X_1)$ for $x \in [-M, M]$, proving Prop. 8. It also proves Prop. 9 by showing that for $x < -M$, $E_x V(X_1) \leq d_{20} + |\theta|\,|x|$, since $\mu = g^{-1}(x) \leq g^{-1}(0)$ on this set.

A.9 Proof of Proposition 10, Case 2

Using Jensen's inequality and the fact that $P_x(Y_0 < c) \overset{x \to \infty}{\longrightarrow} 0$, for all x large enough

$$\begin{aligned} E_x[g(Y_0^*)] \leq g(E_x Y_0^*) &= g(E_x[Y_0 \mathbf{1}_{Y_0 \geq c}] + c P_x(Y_0 < c)) \\ &= g(E_x[Y_0] - E_x[Y_0 \mathbf{1}_{Y_0 < c}] + c P_x(Y_0 < c)). \end{aligned}$$

Using a similar argument to (23) above, we see that the last two terms in the argument of g converge to 0 as $x \to \infty$. Hence, for any $\epsilon > 0$ we can find $M > 0$ so that, for all $x > M$,

$$E_x[g(Y_0^*)] \leq g(g^{-1}(x) + \epsilon) \leq x + d_{21}\epsilon,$$

where d_{21} is the slope of a subgradient of g at $g^{-1}(M)$.

Combining this with (25), there exists $M > 0$ such that for $x > M$,

$$E_x V(X_1) \leq d_{22} + |\rho| V(x) + |\theta| E_x |g(Y_0^*) - x|.$$

It remains to show that the final term in this expression is small relative to the linear (in $V(x)$) term as $x \to \infty$. This follows in almost identical fashion to the proof of this result in Case 1. We omit the details. □

A.10 Proof of Propositions 8-10, Case 3

Assume WLOG that $g(c) = 0$. Since $g(Y_0^*) \in [g(c), g(a-c)]$,

$$\begin{aligned} E_x V(X_1) &= E_x|(1-\rho)\gamma + (\rho+\theta) g(Y_0^*) - \theta x + \sigma Z_0| \\ &\leq (1-\rho)|\gamma| + \sqrt{2\sigma^2/\pi} + |\rho + \theta|\, E_x\,|g(Y_0^*)| + |\theta||x| \\ &\leq d_{23} + |\rho + \theta|\, g(a-c) + |\theta||x|. \end{aligned}$$

Propositions 8, 9, and 10 follow immediately.

A.11 Proof of Proposition 18

By (23),

$$\begin{aligned} E_x|g(Y_0) - x| &= E_x|g(Y_0 \mathbf{1}_{Y_0 > 0}) - x + g(Y_0 \mathbf{1}_{Y_0 < 0})| \\ &\leq E_x|g(Y_0 \mathbf{1}_{Y_0 > 0}) - x| + E_x|g(Y_0 \mathbf{1}_{Y_0 < 0})| \\ &\leq E_x|g(Y_0 \mathbf{1}_{Y_0 > 0}) - x| + a_0 + a_1 E_x[|Y_0| \mathbf{1}_{Y_0 < 0}] \\ &\leq E_x|g(Y_0 \mathbf{1}_{Y_0 > 0}) - x| + d_5 \end{aligned}$$

for $x > M$.

Using (14), for any fixed $\epsilon \in (0,1)$ and $x > M$,

$$\begin{aligned} & E_x\left[|g(Y_0\mathbf{1}_{Y_0>0}) - x|\mathbf{1}_{Y_0\leq(1-\epsilon)\mu}\right] \\ & \leq xP_x(Y_0 \leq (1-\epsilon)\mu) \\ & \leq xP_x(|Y_0 - \mu| > \epsilon\mu) \\ & \leq \frac{x(d_1\mu^r + d_2)}{\epsilon^{2+\delta}\mu^{2+\delta}} \\ & \leq \frac{d_6 x}{\mu^{2+\delta-r}}. \end{aligned} \tag{26}$$

Recall that for $y \geq 0$, $a_0 + a_1 y \geq g(y)$, so that $a_0 + a_1 g^{-1}(y) \geq y$. Hence $\mu = g^{-1}(x) \geq (x - a_0)/a_1$. So (26) is bounded by

$$\frac{d_7 x}{(x-a_0)^{2+\delta-r}}$$

which converges to 0 as $x \to \infty$ and is therefore bounded by d_8 say for $x > M$. It only remains to show that

$$E_x|g(Y_0\mathbf{1}_{\{Y_0>0\}}) - x|\mathbf{1}_{\{Y_0>(1-\epsilon)\mu\}} = E_x|g(Y_0) - x|\mathbf{1}_{\{Y_0>(1-\epsilon)\mu\}}$$

is "small."

Recall that g is concave on $\mathbb{R}^+$ and so has a subgradient at $(1-\epsilon)\mu$, i.e. there exist $b_0(x), b_1(x)$ such that $g(y) \leq b_0(x) + b_1(x)y$ for $y > 0$, with equality at $y = (1-\epsilon)\mu$. The slope of the chord from $(0,0)$ to $((1-\epsilon)\mu, g((1-\epsilon)\mu))$ is greater than or equal to $b_1(x)$, so

$$b_1(x)(1-\epsilon)\mu \leq g((1-\epsilon)\mu) \leq g(\mu) = x. \tag{27}$$

Furthermore, g is concave so $b_1(x)$ is bounded for $x > M$. We now have

$$\begin{aligned} E_x|g(Y_0) - x|\mathbf{1}_{\{Y_0>(1-\epsilon)\mu\}} & \leq b_1(x)E_x|Y_0 - \mu|\mathbf{1}_{\{Y_0>(1-\epsilon)\mu\}} \\ & \leq b_1(x)E_x|Y_0 - \mu| \\ & \leq b_1(x)\left[E_x|Y_0 - \mu|^{2+\delta}\right]^{1/(2+\delta)} \text{ (Jensen)} \\ & \leq b_1(x)(d_1\mu^r + d_2)^{1/(2+\delta)} \\ & \leq b_1(x)(d_9\mu^{r/(2+\delta)} + d_{10}) \text{ (triangle inequality)} \\ & = d_9 b_1(x)\mu^{r/(2+\delta)} + d_{10}b_1(x) \\ & \leq \frac{d_9 x\mu^{r/(2+\delta)}}{(1-\epsilon)\mu} + d_{11} \text{ (from (27))} \\ & \leq d_{12}x\mu^{-(1-r/(2+\delta))} \\ & \leq d_{12}x\left(\frac{x-a_0}{a_1}\right)^{-(1-r/(2+\delta))} \\ & \leq d_{13}x^{r/(2+\delta)}. \end{aligned}$$

proving the result. □

References

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

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

Bougerol, P., and Picard, N. (1992), "Strict stationarity of generalized autoregressive processes," *Annals of Probability*, 20, 1714–1730.

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods*, 2nd edn, New York: Springer-Verlag.

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

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

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

Durbin, J., and Koopman, S. J. (2000), "Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives," *Journal of the Royal Stastistical Society, Series B*, 62, 3–56.

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

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

Fokianos, K., and Tjostheim, D. (2010), Nonlinear Poisson autoregression,. Submitted; available on request from K. Fokianos, http://www2.ucy.ac.cy/~fokianos.

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

Hairer, M. (2008), "Ergodic theory for infinite-dimensional stochastic processes," *Oberwolfach Reports*, 5(4), 2815–2874.

Hairer, M., and Mattingly, J. C. (2006), "Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing," *Annals of Mathematics*, 164, 993–1032.

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

Léon, L. F., and Tsai, C. (1998), "Assessment of model adequacy for Markov regression time series models," *Biometrics*, 54, 1165–1175.

Li, W. K. (1994), "Time series models based on Generalized Linear Models: Some further results," *Biometrics*, 50, 506–511.

Matteson, D. S., McLean, M. W., Woodard, D. B., and Henderson, S. G. (2011), "Forecasting Emergency Medical Service call arrival rates," *Annals of Applied Statistics*, . In press.

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

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

Roberts, G. O., and Rosenthal, J. S. (2004), "General state space Markov chains and MCMC algorithms," *Probability Surveys*, 1, 20–71.

Talamantes, J., Behseta, S., and Zender, C. S. (2007), "Statistical modeling of valley fever data in Kern County, California," *International Journal of Biometeorology*, 51, 307–313.

Thorisson, H. (1995), "Coupling methods in probability theory," *Scandinavian Journal of Statistics*, 22, 159–182.

Tweedie, R. L. (1988), "Invariant measures for Markov chains with no irreducibility assumptions," *Journal of Applied Probability*, 25, 275–285.

Zeger, S. L. (1988), "A regression model for time series of counts," *Biometrika*, 75, 621–629.

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