Stationarity of Count-Valued and Nonlinear Time Series Models

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

October 21, 2010

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. One of the most general methods for proving stationarity is via the use of drift conditions; however, this method assumes φ-irreducibility, which is violated by the important class of count-valued observation-driven models. We provide a formal justification for the use of drift conditions on count-valued observation-driven models, and demonstrate by proving for the first time stationarity and ergodicity of several models. These include the class of Generalized Autoregressive Moving Average models, which contains a number of important count-valued and nonlinear models as special cases.

Keywords: ergodicity; drift conditions; stationary; irreducibility; autoregressive moving average model.

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 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.

One 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. Such a strong statement of stationarity is quite useful, since it immediately implies 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, Lyapunov function methods assume φ-irreducibility, which is violated by count-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-valued time series. We provide a formal justification for the use of drift conditions on count-valued observation-driven models. In particular, we introduce small random innovations to the process that induce φ-irreducibility, and we prove the validity of obtaining stationarity and ergodicity results for the perturbed instead of original process. Our results apply to the entire class of observation-driven models. A similar perturbation method for a specific example is used in Fokianos et al. (2009); our results for our perturbation approach are much more general, and our justification for analyzing the perturbed process is unrelated to that of Fokianos et al. (2009).

We demonstrate our approach by proving for the first time stationarity and ergodicity of several important models, including Generalized Autoregressive Moving Average (GARMA) models (Benjamin et al., 2003). GARMA models generalize autoregressive moving average models to exponential-family distributions, naturally handling count- and positive-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 and ergodicity 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).

Lyapunov function methods have been previously applied to prove stationarity and ergodicity for a variety of real-valued time series models, including for SETAR models by Chan and Tong (1985), for multivariate ARMA models by Bougerol and Picard (1992), for threshold AR-ARCH models by Cline and Pu (2004), and for integrated GARCH models by Liu (2009). Most of these papers use the Lyapunov exponent, which can give a sufficient and necessary condition for stationarity (Bougerol and Picard, 1992; Cline and Pu, 2004; Liu, 2009); however, this condition has a complex form that can be difficult to reduce to simple ranges on the parameter values. We instead take the constructive approach used, e.g., in Chan and Tong (1985), which is much more straightforward to understand and apply. It gives sufficient but not necessary conditions for stationarity; however, these conditions often have a much simpler form than those given by the Lyapunov exponent.

In Section 2 we describe Lyapunov function methods and give our justification for using these methods on count-valued time series models. In Section 3 we illustrate by application to a specific linear count-valued model, and in Section 4 we use our method to prove stationarity for the class of GARMA models.

2 Drift Conditions for Observation-Driven Models

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

$$Y_n | Y_{0:n-1} \overset{\text{ind}}{\sim} \psi_\nu(\mu_n) \tag{1}$$

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

for some function $h_{\theta,n}$ 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 best 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.

Next 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 $x \in S$,*

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

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

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

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 value $\mu_{0:p-1}$ the set of possible values for μ_n is countable. In fact, the set of states that are reachable by the Markov chain $\{\mu_{n-p:n-1}\}_{n\geq p}$ from a fixed starting state is also countable, and distinct initial values can have distinct sets of reachable locations. 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 are reachable from any fixed X_0 is countable. Chains with these kinds of properties have been studied (Borovkov, 1998), but require much more technical care and model-specific arguments than the method we present here.

We will show that one can instead prove stationarity of a slightly perturbed model (the perturbations are used only for purposes of analysis, not when applying the model). We do this by returning to the most general framework (1) and (2), and replacing $h_{\theta,n}$ with a function of two inputs:

$$\mu_n^{(\sigma)} = h_{\theta,n}(Y_{0:n-1}, \sigma Z_{0:n-1}) \tag{6}$$

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 $h_{\theta,n}(Y_{0:n-1}, 0) = h_{\theta,n}(Y_{0:n-1})$ for any $Y_{0:n-1}$. When the perturbed model is constructed to be φ-irreducible, one can then apply drift conditions to prove its stationarity.

We will show that there is no practical impact of proving stationarity for the perturbed rather than the original model because, loosely speaking, the likelihood of the parameter vector $\eta = (\theta, \nu)$ calculated using (6) converges to the likelihood calculated using (2) as $\sigma \to 0$. More precisely, the joint density of the observations $Y = Y_{0:N}$ and first N perturbations $Z = Z_{0:N-1}$, conditional on the parameter vector η and the perturbation scale σ is

$$\begin{aligned} f(Y, Z|\eta, \sigma) &= f(Z|\eta, \sigma) \times f(Y|Z, \eta, \sigma) \\ &= \left[\prod_{n=0}^{N-1} \phi(Z_n)\right] \prod_{n=0}^{N} \psi_\nu(Y_n; \mu_n(\sigma Z)) \end{aligned}$$

where $\mu_n(\sigma Z)$ is the value of $\mu_n^{(\sigma)}$ induced by the perturbation vector σZ through (6), with $\mu_0(\sigma Z) = \mu_0$ being a fixed initial value. The likelihood function for the parameter vector η implied by the perturbed model is the marginal density of Y integrating over Z, i.e.,

$$\mathcal{L}_\sigma(\eta) = f(Y|\eta, \sigma) = E\left[\prod_{n=0}^{N} \psi_\nu(Y_n; \mu_n(\sigma Z)) \,\middle|\, Y\right].$$

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

$$\mathcal{L}(\eta) = \prod_{n=0}^{N} \psi_\nu(Y_n; \mu_n(0)).$$

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.,*

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

for any fixed sequence of observations $Y = Y_{0:N}$. 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).

Proof. Fixing $Y_{0:N}$ and letting $Z = Z_{0:N-1}$ be the perturbations,

$$
\begin{aligned}
\sup_{\eta\in K} |\mathcal{L}_\sigma(\eta) - \mathcal{L}(\eta)| &= \sup_{\eta\in K} \left| E \prod_{n=0}^{N} \psi_\nu(Y_n;\mu_n(\sigma Z)) - \prod_{n=0}^{N} \psi_\nu(Y_n;\mu_n(0)) \right| \\
&\le \sup_{\eta\in K} E \left| \prod_{n=0}^{N} \psi_\nu(Y_n;\mu_n(\sigma Z)) - \prod_{n=0}^{N} \psi_\nu(Y_n;\mu_n(0)) \right| \\
&\le E \sup_{\eta\in K} \left| \prod_{n=0}^{N} \psi_\nu(Y_n;\mu_n(\sigma Z)) - \prod_{n=0}^{N} \psi_\nu(Y_n;\mu_n(0)) \right| \\
&= E \sup_{\eta\in K} \left| \prod_{n=0}^{N} \beta_n(\sigma Z) - \prod_{n=0}^{N} \beta_n(0) \right| \qquad (7)
\end{aligned}
$$

where $\beta_n(\cdot) = \psi_\nu(Y_n;\mu_n(\cdot))$. We will show that the supremum inside the expectation in (7) converges to 0 almost surely (in Z) as $\sigma \to 0$; then bounded convergence implies that the expectation (7) 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 $\mu_n(\cdot)$ is Lipschitz continuous in some bounded neighborhood C of 0, uniformly in $\eta \in K$. In other words, there exists a finite constant L_n such that, for any $z, z' \in C$,

$$
\sup_{\eta\in K} |\mu_n(z) - \mu_n(z')| \le L_n \|z - z'\|
$$

for each $n = 0, 1, \ldots, N$. Thus, the composition $\beta_n(\cdot) = \psi_\nu(Y_n, \mu_n(\cdot))$ is Lipschitz continuous on C, uniformly in $\eta \in K$, for each $n = 0, 1, \ldots, N$.

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

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

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

$\square$

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., 2010). The model is defined as

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

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 $\eta \le 0$ and $\gamma \ge 0$, 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 $\{Y_n\}_{n\in\mathbb{N}}$ is stationarity 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, where $X_n = \mu_n$. First, incorporate perturbations $Z_n \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$ as in Theorem 2:

$$\mu_n = \omega + \alpha Y_{n-1} + \beta\mu_{n-1} + (\gamma Y_{n-1} + \eta\mu_{n-1})\,\mathbf{1}_{\{Y_{n-1}\notin(L,U)\}} + \sigma Z_{n-1}.$$

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

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 $V(x) = x$,

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

In particular, $E_x V(X_1)$ is bounded for $x \in A$. Also, as $x \to \infty$ we have $x P_x[Y_0 \in (L,U)] \to 0$, so for sufficiently large M, $x > M$ implies that $-\eta x P_x[Y_0 \in (L,U)] \le \frac{\sigma}{2}$. Thus for $x > M$,

$$E_x V(X_1) \;\; \le \;\; (\alpha+\beta+\gamma+\eta)V(x) + (\omega+\sigma) \le \nu V(x)$$

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 $V^*(x) = x + 1$ to get the range $[1,\infty)$. So the chain X is geometrically ergodic, and thus stationary for an appropriate initial distribution. As shown in Section 2, this implies that the time series model $\{Y_n\}_{n\in\mathbb{N}}$ is also stationary and ergodic.

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 positive 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 very general form:

$$Y_n|\mathbf{D}_n \overset{\text{ind}}{\sim} \psi_\nu(\mu_n) \tag{8}$$

$$g(\mu_n) = W_n'\beta + \sum_{j=1}^{p} \rho_j[g(Y_{n-j}^*) - W_{n-j}'\beta] + \sum_{j=1}^{q} \theta_j[g(Y_{n-j}^*) - g(\mu_{n-j})] \tag{9}$$

for some real-valued link function g, where Y_n^* is some modification of Y_n that maps it to the domain of g, where W_n are the covariates at time n, and where

$$\mathbf{D}_n = (W_{n-p}, \ldots, W_n, Y_{n-\max\{p,q\}}, \ldots, Y_{n-1}, \mu_{n-q}, \ldots, \mu_{n-1})$$

are the present covariates and the relevant past information (Benjamin et al., 2003). The second and third terms of the model (9) are the autoregressive and moving-average terms, respectively. This model is more general than the class of models developed in Benjamin et al. (2003) because we do not assume that ψ_ν is in the exponential family. However, we do assume that $E(Y_n|\mathbf{D}_n) = \mu_n$, 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 normal, Poisson, and binomial GARMA models.

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

$$g(\mu) = \log\left(e^{\alpha\mu} - 1\right)/\alpha \tag{10}$$

which has the property $g(\mu) \approx \mu$ 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.

Since the covariates are time-dependent, the model for $\{Y_n\}_{n\in\mathbb{N}}$ is in general nonstationary, and interest is in proving stationarity in the absence of covariates (i.e. for the case $W_n = 1$). For simplicity we prove the case $p = 1$ and $q = 1$ here; the extension to $p > 1$ and $q > 1$ is sketched at the end of Sec. 4.1. Let $\rho = \rho_1$ and $\theta = \theta_1$, and denote the intercept by γ, yielding the perturbed model:

$$g(\mu_n) = \gamma + \rho[g(Y^*_{n-1}) - \gamma] + \theta[g(Y^*_{n-1}) - g(\mu_{n-1})] + \sigma Z_{n-1} \tag{11}$$

where $Z_n \overset{\text{iid}}{\sim} N(0, 1)$, for any $\sigma > 0$.

For this model, we have the following stationarity results:

Theorem 3. *The process $\{Y_n\}_{n\in\mathbb{N}}$ specified by the GARMA model (8) and (11) is an ergodic Markov chain and thus stationary for an appropriate initial distribution for μ_0, provided that:*

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

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

The following popular models are special cases of Theorem 3:

Corollary 4. *Suppose that conditional on μ_n, Y_n is normally distributed with mean μ_n and fixed variance $\tau^2 > b$ that is restricted from below by a known constant $b > 0$. Then the GARMA model is ergodic and thus stationary for an appropriate initial distribution for μ_0, provided that $|\rho| < 1$ and the link function is bijective, increasing, concave on $\mathbb{R}^+$, and convex on $\mathbb{R}^-$. For instance, this is satisfied by the identity link or the symmetric power link*

$$g(\mu) = \begin{cases} \mu^\alpha & \mu \geq 0 \\ -|\mu|^\alpha & \mu < 0. \end{cases}$$

for $\alpha \in (0, 1)$.

The lower bound on τ^2 ensures that the regularity conditions for Theorem 2 are satisfied.

Proof. A normal random variable X with mean μ and variance τ^2 satisfies

$$E(X - \mu)^4 = 3\tau^4.$$

So in our case we can take $\delta = 2$ and $r = 0$. Theorem 2 applies here, shown as follows. The normal density ψ_ν satisfies regularity condition (a) when $\tau^2 > b$. Also, $X_n = g(\mu_n)$ is linear in $Z_{0:n-1}$ and $g^{-1}(\cdot)$ is Lipschitz on any compact set (due to the concavity/convexity restrictions on 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 $(\gamma, \rho, \theta) \in \mathbb{R}^3$. □

The classical ARMA model with $p, q = 1$ is an instance of Corollary 4 with identity link function. In this case the parameter restriction $|\rho| < 1$ is well-known to be necessary and sufficient for stationarity.

Corollary 5. *Suppose that conditional on μ_n, Y_n is Poisson distributed with mean μ_n. Then the GARMA model is ergodic and thus stationary for an appropriate initial distribution for μ_0, 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 (10).*

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, by verifying the regularity conditions as for Corr. 4. □

Corollary 6. *Suppose that conditional on μ_n, Y_n is binomially distributed with mean μ_n and fixed number of trials k. Then the GARMA model is ergodic and thus stationary for an appropriate initial distribution for μ_0, provided that $|\theta| < 1$, g is bijective (e.g. the logit link), and g^{-1} is Lipschitz on any compact set.*

The additional condition on g^{-1} ensures that Theorem 2 applies. This condition 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 $r = 0$:

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

Theorem 2 applies here, by verifying the regularity conditions as for Corr. 4. Unlike the case of Corr. 4, g^{-1} is not automatically Lipschitz on any compact set, which is why Corr. 6 explicitly makes this assumption. □

4.1 Proof of Theorem 3

Define $X_n = g(\mu_n)$; we will prove Theorem 3 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

$$P_x(|Y_0 - \mu| > y) \leq \frac{E_x|Y_0-\mu|^{2+\delta}}{y^{2+\delta}} \leq \frac{d_1|\mu|^r + d_2}{y^{2+\delta}}. \tag{12}$$

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

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

Then with probability at least 3/4,

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

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 7. ***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 8. ***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 9. ***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 and 9 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. 9. 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 $V^*(x) = |x| + 1$ to get the range $[1, \infty)$).

These results have the following intuition for Case 2: Prop. 8 shows that for very negative X_{n-1}, $|\theta|$ controls the rate of drift, while Prop. 9 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 (11) 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} concentrates around μ_{n-1}, so that the moving-average term $\theta[g(Y^*_{n-1}) - g(\mu_{n-1})]$ in (11) is negligible and the autoregressive term dominates.

Extension to the cases $p > 1$ and $q > 1$ can be achieved by showing geometric ergodicity of the multivariate Markov chain with state vector $\mu_{(n-\max\{p,q\}+1):n}$. Again this is 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 $(\theta_1, \ldots, \theta_q)$.

4.2 Proof of Prop. 7, Case 1

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 (11), and the triangle inequality, we can bound $E_x V(X_1)$ as follows, where d_i denote bounded (in μ) constants for each $i \geq 3$:

$$\begin{aligned} E_x V(X_1) &= E_x|(1-\rho)\gamma + \rho g(Y_0) + \theta(g(Y_0) - x) + \sigma Z_0| \\ &\leq (1-\rho)|\gamma| + \sqrt{2\sigma^2/\pi} + |\rho| E_x|g(Y_0)| + |\theta| E_x|g(Y_0) - x| \\ &\leq d_3 + (|\rho| + |\theta|)a_1 E_x|Y_0| + |\theta||x|. \end{aligned} \tag{13}$$

By the triangle and Jensen's inequalities,

$$\begin{aligned} E_x|Y_0| &= E_x|\mu + Y_0 - \mu| \\ &\leq |\mu| + E_x|Y_0 - \mu| \\ &\leq |\mu| + \left[E_x|Y_0 - \mu|^{2+\delta}\right]^{1/(2+\delta)} \\ &\leq |\mu| + (d_1|\mu|^r + d_2)^{1/(2+\delta)}. \end{aligned} \tag{14}$$

So $\sup_{x\in[-M,M]} E_x V(X_1) < \infty$, proving Prop. 7.

4.3 Proof of Prop.s 8 and 9, Case 1

We will prove Prop. 9 for Case 1; Prop. 8 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 (13), 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

$$\begin{aligned} E_x|g(Y_0)| &= E_x[g(Y_0)\mathbf{1}_{Y_0>0}] - E_x[g(Y_0)\mathbf{1}_{Y_0<0}] \\ &= E_x g(Y_0\mathbf{1}_{Y_0>0}) - E_x g(Y_0\mathbf{1}_{Y_0<0}) \\ &\leq g(E_x[Y_0\mathbf{1}_{Y_0>0}]) - g(E_x[Y_0\mathbf{1}_{Y_0<0}]) \\ &= g(E_x Y_0 - E_x[Y_0\mathbf{1}_{Y_0<0}]) - g(E_x[Y_0\mathbf{1}_{Y_0<0}]) \end{aligned} \tag{15}$$

by Jensen's inequality. Now, $\mu = g^{-1}(x) > 0$ for $x > 0$, so using (12)

$$\begin{aligned} -E_x[Y_0\mathbf{1}_{Y_0<0}] &= \int_0^\infty P_x(Y_0 < -u)\,du \\ &\leq \int_0^\infty P_x(|Y_0 - \mu| > u + \mu)\,du \\ &\leq \int_0^\infty \frac{d_1\mu^r + d_2}{(u+\mu)^{2+\delta}}\,du \\ &= \frac{d_1\mu^r + d_2}{(1+\delta)\mu^{1+\delta}} \to 0 \end{aligned} \tag{16}$$

as $x \to \infty$. Thus, from (15), for any given $\epsilon > 0$, there exists $M > 0$ so that for $x > M$,

$$E_x|g(Y_0)| \leq g(E_x Y_0 + \epsilon) + \epsilon \leq g(E_x Y_0) + g(\epsilon) + \epsilon = x + d_4 \tag{17}$$

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 (13) is "small" relative to the linear (in x) term:

Proposition 10. *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. 10 is proven in the Appendix. Combining it with (13) and (17), we have that for all x large enough,

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

proving Prop. 9. □

4.4 Proof of Prop. 7 and Prop. 8, 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 (11), and the triangle inequality, we can bound $E_x V(X_1)$ as follows:

$$\begin{aligned} E_x V(X_1) &= E_x|(1-\rho)\gamma + \rho g(Y_0^*) + \theta(g(Y_0^*) - x) + \sigma Z_0| \\ &\leq (1-\rho)|\gamma| + \sqrt{2\sigma^2/\pi} + |\rho| E_x[g(Y_0^*)] + |\theta| E_x|g(Y_0^*) - x| \qquad (18) \\ &= d_{15} + |\rho| P_x(Y_0 < c) g(c) + |\rho| E_x[g(Y_0)\mathbf{1}_{Y_0 \geq c}] + \\ &\quad |\theta| P_x(Y_0 < c)|g(c) - x| + |\theta| E_x[|g(Y_0) - x|\mathbf{1}_{Y_0 \geq c}] \\ &\leq d_{15} + (|\rho| + |\theta|) E_x[g(Y_0)\mathbf{1}_{Y_0 \geq c}] + \\ &\quad |\theta| P_x(Y_0 < c)|g(c) - x| + |\theta| P_x(Y_0 \geq c)|x| \\ &\leq d_{15} + (|\rho| + |\theta|) a_1 E_x[Y_0 \mathbf{1}_{Y_0 \geq c}] + |\theta||x| \end{aligned}$$

In the same way that we obtained (14) 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. 7. It also proves Prop. 8 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.

4.5 Proof of Prop. 9, 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

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

Using a similar argument to (16) 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 $x > M$,

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

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

Combining this with (18), there exists $M > 0$ such that for $x > M$,

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

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. $\square$

4.6 Proof of Prop.s 7-9, Case 3

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

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

Propositions 7, 8, and 9 follow immediately.

Acknowledgments

This work was supported in part by National Science Foundation Grant Number CMMI-0926814.

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.

Borovkov, A. A. (1998), *Ergodicity and Stability of Stochastic Processes*, 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.

Chan, K. S., and Tong, H. (1985), "On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations," *Advances in Applied Probability*, 17, 666–678.

Cline, D. B. H., and Pu, H. H. (2004), "Stability and the Lyapunov exponent of threshold AR-ARCH models," *Annals of Applied Probability*, 14, 1920–1949.

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.

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.

Liu, J. (2009), "Stationarity of a family of GARCH processes," *Econometrics Journal*, 12, 436–446.

Matteson, D. S., Woodard, D. B., and Henderson, S. G. (2010), Forecasting Emergency Medical Service call arrival rates,. Submitted.

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.

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.

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.

Appendix: Proof of Prop. 10

By (16),

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

for $x > M$.

Using (12), for any fixed $\epsilon \in (0,1)$ and $x > M$,

$$\begin{aligned}
& E_x\left[|g(Y_0\mathbf{1}_{Y_0>0}) - x|\mathbf{1}_{Y_0\le(1-\epsilon)\mu}\right] && (19)\\
& \le xP_x(Y_0 \le (1-\epsilon)\mu) \\
& \le xP_x(|Y_0-\mu| > \epsilon\mu) \\
& \le \frac{x(d_1\mu^r + d_2)}{\epsilon^{2+\delta}\mu^{2+\delta}} \\
& \le \frac{d_6 x}{\mu^{2+\delta-r}}.
\end{aligned}$$

Recall that for $y \ge 0$, $a_0 + a_1 y \ge g(y)$, so that $a_0 + a_1 g^{-1}(y) \ge y$. Hence $\mu = g^{-1}(x) \ge (x-a_0)/a_1$. So (19) 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 $g(y) \le b_0(x) + b_1(x)y$ 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

$$b_1(x)(1-\epsilon)\mu \le g((1-\epsilon)\mu) \le g(\mu) = x. \qquad (20)$$

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

$$\begin{aligned}
E_x|g(Y_0) - x|\mathbf{1}_{\{Y_0>(1-\epsilon)\mu\}} &\le b_1(x)E_x|Y_0 - \mu|\mathbf{1}_{\{Y_0>(1-\epsilon)\mu\}} \\
&\le b_1(x)E_x|Y_0 - \mu| \\
&\le b_1(x)\left[E_x|Y_0 - \mu|^{2+\delta}\right]^{1/(2+\delta)} \quad \text{(Jensen)} \\
&\le b_1(x)(d_1\mu^r + d_2)^{1/(2+\delta)} \\
&\le b_1(x)(d_9\mu^{r/(2+\delta)} + d_{10}) \text{ (triangle inequality)} \\
&= d_9 b_1(x)\mu^{r/(2+\delta)} + d_{10}b_1(x) \\
&\le \frac{d_9 x\mu^{r/(2+\delta)}}{(1-\epsilon)\mu} + d_{11} \text{ (from (20))} \\
&\le d_{12}x\mu^{-(1-r/(2+\delta))} \\
&\le d_{12}x\left(\frac{x - a_0}{a_1}\right)^{-(1-r/(2+\delta))} \\
&\le d_{13}x^{r/(2+\delta)}.
\end{aligned}$$

proving the result. □