Electronic Journal of Statistics
Vol. 16 (2022) 1393–1433
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS1989

Observation-driven models for discrete-valued time series*

Mirko Armillotta

Department of Mathematics & Statistics, University of Cyprus,
PO BOX 20537, Nicosia, Cyprus
e-mail: armillotta.mirko@ucy.ac.cy

Alessandra Luati

Department of Statistical Sciences, University of Bologna,
41 st. Belle Arti, 40126, Bologna, Italy
e-mail: alessandra.luati@unibo.it

Monia Lupparelli

Department of Statistics, Computer Science, Applications, University of Florence,
59 st. Morgagni, 50134, Florence, Italy
e-mail: monia.lupparelli@unifi.it

Abstract: Statistical inference for discrete-valued time series has not been developed like traditional methods for time series generated by continuous random variables. Some relevant models exist, but the lack of a homogenous framework raises some critical issues. For instance, it is not trivial to explore whether models are nested and it is quite arduous to derive stochastic properties which simultaneously hold across different specifications. In this paper, inference for a general class of first order observation-driven models for discrete-valued processes is developed. Stochastic properties such as stationarity and ergodicity are derived under easy-to-check conditions, which can be directly applied to all the models encompassed in the class and for every distribution which satisfies mild moment conditions. Consistency and asymptotic normality of quasi-maximum likelihood estimators are established, with the focus on the exponential family. Finite sample properties and the use of information criteria for model selection are investigated throughout Monte Carlo studies. An empirical application to count data is discussed, concerning a test-bed time series on the spread of an infection.

MSC2020 subject classifications: Primary 62M20, 62F12; secondary 62M10, 62J12.

Keywords and phrases: Count data, generalized ARMA models, likelihood inference, link-function.

Received November 2021.

*This work has been co-financed by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation, under the project INFRASTRUCTURES/1216/0017 (IRIDA).

1. Introduction

The analysis of time series that are generated by continuous random variables has a long tradition in statistics and dates back, in the parametric setting, to [42] and [41], who introduced the concept of autoregression, a dynamic model for the conditional mean of a stochastic process. In the same period, [37] defined moving average processes as linear combinations of uncorrelated random variables capable of capturing cyclical fluctuations. It was only in the seventies, with the formalization by [4] of the class of ARMA models, that autoregressive (AR) and moving average (MA) processes found their popularity and became massively fitted to real data. The merit of Box and Jenkins's work was the specification of a unified class of processes, generalizing ARMA models to account for non-stationarity, seasonality, exogenous regressors, as well as the systematic treatment of all the sub-models belonging to the class, which led to the development of well established inferential procedures.

The development of parametric models for count and binary data has not enjoyed the same popularity, partly because linear processes are related to second order stationarity, which fully characterizes Gaussian time series. For discrete data, the concept of autocovariance needs to be adapted [38] and the Wold representation has no direct interpretation, see the discussion in the recent handbook edited by [11]. Since the AR- and MA-like models first introduced by [43] and [27], there have been some relevant specifications, such as the generalized ARMA (GARMA) by [3] and their martingalized version, the M-GARMA by [44], as well as the generalized linear ARMA (GLARMA) by [9]. An interesting class of autoregression models for count data has been proposed by [20] and [22], inspired by the generalized linear transformation of [30]. Integer-valued time series with extreme observations have been dealt with by [25], based on the beta-Negative Binomial distribution.

As recently acknowledged by [10], the analysis of discrete-valued time series would benefit from the specification of a unified framework able to encompass most of the models available in the literature. As a matter of fact, it is not trivial to explore whether models are nested, and, consequently, to derive stochastic properties that simultaneously hold across models. In addition, model comparison becomes crucial when direct relationships among different models are unknown. The lack of a unified framework and, consequently, of systematic analysis is also in contrast with the growing attention paid, in recent years, to high dimensional data sets involving dynamic binary and count data, in different contexts, such as number of clicks or amount of intra-day stock transactions [12, 1]. Attempts in this direction have been made by [14] who provide a theoretical formulation which is useful in principle but less effective when the aim is to implement and adapt models for real applications. Indeed, the quite general framework developed by [14] encompasses several models for which stochastic and inferential properties have been previously derived in the literature, but at the price of conditions that are extremely complicated to verify in practice for each model and distribution.

To summarize the main results developed in the literature, on the side of the

stochastic properties, [29] develop notable results about strict stationarity and ergodicity for the specific case of GARMA and Poisson Threshold autoregressive models, using the theory of Markov chains. Conversely, conditions holding for several models but requiring restrictive assumptions are discussed in [31], based on contraction conditions, and in [16], based on the weak dependence approach. [20] and [22] develop results on ergodicity employing a perturbation approach, which is necessarily suited for the case of count data following a Poisson distribution. Similar results are discussed in [5] under the assumption of a Negative Binomial distribution as the data generating process. An interesting extension to categorical time series has been recently proposed by [23].

As far as inference is concerned, the properties of the maximum likelihood estimator (MLE) and Quasi MLE (QMLE) have been studied for some subsets of discrete-valued models. [14] prove the consistency of MLE and QMLE for the general framework they propose. Asymptotic normality, in the same setting, is later discussed by [15]. Comparable results have been derived by [12], based on the approach developed by [31], and by [1] for the specific case of the Poisson distribution. However, the conditions needed to verify the properties of MLE and QMLE are far from immediate.

This paper introduces a general class of observation-driven models for discrete-valued stochastic processes that encompasses the existing models in the literature and includes novel specifications. In the terminology of [7], observation-driven models are designed for time-varying parameters whose dynamics are functions of the past observations only and are not driven by an idiosyncratic noise term. Essentially, we specify a dynamic model for the conditional mean of a density, or mass function for discrete-valued time series, which does not necessarily belong to the exponential family. This generality allows one to estimate alternative models designed to capture the past effects of the conditional mean itself, of the lagged discrete-valued process and error-type components.

The stochastic theory and the likelihood inference are developed for first order models in the class (i.e. with one lag of autoregression), through an extension of the theory of [29] as far as stationarity and ergodicity are concerned, and of [14] and [15] for the asymptotic properties of likelihood estimators. In addition to the results that apply to novel models, we derive several new methodological results for existing models, that were not yet proved in the literature, such as strict-stationarity and ergodicity of first order GLARMA models and ergodicity of M-GARMA models for discrete distributions.

In summary, a general modelling framework is introduced which aims (i) to provide a unified specification for a broad class of discrete-valued time series where relevant instances represent special cases, (ii) to provide direct relationships among different models which belong to the framework but are not necessarily nested within each other, (iii) to derive the stochastic properties for first order models which hold simultaneously for the entire class (strict stationarity and ergodicity), (iv) to implement QMLE inference that also allows us to define model selection criteria across different, and not nested, models, (v) to derive the asymptotic properties of QMLE, and (vi) to make all the models encompassed in the framework fully applicable in practice. With the focus on model

comparison, models included in the general framework are applied for the analysis of a test-bed time series in count data analysis, on the spread of an infection, namely Escherichia coli, in the German region of North Rhine-Westphalia.

2. The general framework

Let $\{Y_t\}_{t\in T}$ be a stationary stochastic process defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ where $\mathcal{F} = \{\mathcal{F}_t\}_{t\in T}$ and $\mathcal{F}_t = \sigma(Y_{t-s}, s \geq 0)$ is the sigma-algebra generated by the random variables Y_s, $s \leq t$, with $\mathrm{E}|Y_t| < \infty$ for all $t \in T$. We specify a class of observation-driven models where the conditional density or mass function of Y_t, depending on a time-varying parameter $\mu_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$, is a member of the one-parameter exponential family

$$q(Y_t|\mathcal{F}_{t-1}) = \exp\left\{Y_t\, f(X_t) - A\left(X_t\right) + d(Y_t)\right\}\,, \tag{1}$$

$$X_t = g(\mu_t) = \mathbf{Z}_t^T\boldsymbol{\alpha} + \sum_{j=1}^{k}\gamma_j g(\mu_{t-j}) + \sum_{j=1}^{p}\phi_j h(Y_{t-j}) + \sum_{j=1}^{q}\theta_j\left[\frac{h(Y_{t-j}) - \bar{g}(\mu_{t-j})}{\nu_{t-j}}\right], \tag{2}$$

where the dynamics of the density (or mass) function $q(Y_t|\mathcal{F}_{t-1})$ are captured by the parameter μ_t, or equivalently by X_t. The time-varying parameter μ_t is related to the process X_t by a twice-differentiable, one-to-one monotonic function $g(\cdot)$, which is called the link function. The functions $A(\cdot)$ (log-partition) and $d(\cdot)$ define the particular distribution [30]. The mapping $f(\cdot)$ is a twice-differentiable bijective function, chosen according to the model of interest. Note that the process $\{Y_t\}_{t\in T}$ is observed whereas $\{\mu_t\}_{t\in T}$ is not. However, from equation (2), it can be shown, by backward substitutions, that the process $\{\mu_t\}_{t\in T}$ is a deterministic function of the past $\mathcal{F}_{t-1}$ and this is the reason why we refer to "observation-driven models". The function $h(Y_t)$ is called the "data-link function" since it is applied to the process Y_t whereas $\bar{g}(\mu_t)$ is said the "mean-link function" since it is applied only to the conditional mean, unlike the link function $g(\cdot)$ which, in principle, can be applied to any parameter or moment of the probability distribution. Both the functions $h(Y_t)$ and $\bar{g}(\mu_t)$ are twice-differentiable, one-to-one monotonic and their shape depends on the specific model (2) and the distribution of interest in equation (1).

The vector $\mathbf{Z}_t = [1, Z_{1t}, \ldots, Z_{st}]^T$ in equation (2) is a vector of covariates and $\boldsymbol{\alpha}$ is the corresponding coefficient vector with comparable dimensions. The parameters ϕ_j measure an autoregressive-like effect of the observations; instead, the parameters γ_j state the dependence of the process from its whole past memory (since μ_{t-j} depends on the past observations $Y_{t-j-1}, \ldots$); finally, the parameters θ_j represent the analogous of a moving average component, since the last term of (2) can be defined as a prediction error

$$\varepsilon_t = \frac{h(Y_t) - \bar{g}(\mu_t)}{\nu_t}\,, \tag{3}$$

where the process $\{\nu_t\}_{t\in T}$ is some scaling sequence, typically: (*i*) $\nu_t = \sigma_t$ (Pearson residuals), (*ii*) $\nu_t = \sigma_t^2$ (score-type residuals), (*iii*) $\nu_t = 1$ (no scaling),

and (iv) $\nu_t = \sqrt{\mathrm{V}[h(Y_t)\,|\mathcal{F}_{t-1}]}$. Note that every time the mean-link function is selected as the conditional expectation of the data-link function for the process, $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]$, the difference $h(Y_t) - \bar{g}(\mu_t)$ is a martingale difference sequence (MDS). Moreover, if $\nu_t = \sqrt{\mathrm{V}[h(Y_t)\,|\mathcal{F}_{t-1}]}$, then the residuals in equation (3) form a white noise (WN) sequence, with unit variance.

Each exponential family in the form (1) can be re-parametrized in the canonical form:

$$q(Y_t|\mathcal{F}_{t-1}) = \exp\left\{Y_t\, Q_t - \bar{A}\,(Q_t) + d(Y_t)\right\}\,, \tag{4}$$

where the sequence $Q_t = f(X_t) = f[g(\mu_t)] = \tilde{f}(\mu_t)$ is called the canonical parameter, whereas the function $\tilde{f}(\cdot) = (f \circ g)(\cdot)$ is referred to as the canonical link function and $\bar{A}\,(\cdot)$ is a re-parametrization of $A\,(\cdot)$ with respect to Q_t. It is known that for the exponential family (4) the conditional mean is $\mu_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1}) = \bar{A}'(Q_t) = \tilde{f}^{-1}(Q_t) = g^{-1}(X_t)$ and the conditional variance is $\sigma_t^2 = \mathrm{V}(Y_t\,|\mathcal{F}_{t-1}) = \bar{A}''(Q_t)$. If $g(\cdot)$ is the canonical link function, then $\tilde{f} \equiv g$ and the following simplification occurs: $f(X_t) = X_t$, so $Q_t = X_t = g(\mu_t)$, which gives again the distribution (1), with $f(X_t) = X_t$, so that (1) and (4) are exactly the same. Clearly, the moments become $\mu_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1}) = A'(X_t) = g^{-1}(X_t)$ and $\sigma_t^2 = \mathrm{V}(Y_t\,|\mathcal{F}_{t-1}) = A''(X_t)$. The function $f(\cdot)$ allows us to introduce non-canonical shapes for $g(\cdot)$, thus adding flexibility to the model. We provide some examples to clarify the nature of the framework.

Example 1 In the setting (1, 2), the Poisson distribution is obtained with $f(X_t) = X_t$, $g(\mu_t) = \log(\mu_t)$, $A\,[g(\mu_t)] = \mu_t$ and $d(Y_t) = \log(1/Y_t!)$. All the derivatives of $A(X_t) = \exp(X_t)$ equal μ_t. However, this definition is based on the equivalence $g \equiv \tilde{f}$, which is the canonical link; hence equation (2) becomes a log-linear model on the response $\log(\mu_t)$. It is possible to model (2) with a different shape of $g(\cdot)$; for example, one may be interested in a linear model for the parameter of the Poisson μ_t, then $g(\mu_t) = \mu_t$ and clearly $g \neq \tilde{f}$. In this case, the Poisson distribution is reconstructed from (1), by setting $f(X_t) = \log(X_t) = \log(\mu_t)$, $A(X_t) = X_t = \mu_t$ and $d(Y_t) = \log(1/Y_t!)$. Again, by knowing that the inverse of the canonical link $\tilde{f}^{-1}(\cdot) = \exp(\cdot)$, the conditional expectation would be $\mathrm{E}(Y_t|\mathcal{F}_{t-1}) = \mathrm{V}(Y_t|\mathcal{F}_{t-1}) = \tilde{f}^{-1}(Q_t) = \exp[f(X_t)] = \mu_t$.

Example 2 The Gaussian distribution (with known variance σ^2) is obtained by setting $d(Y_t) = \log\left[-1/\sqrt{2\pi\sigma^2}\exp\left(-Y_t^2/2\sigma^2\right)\right]$, $f(X_t) = X_t$, $g(\mu_t) = \mu_t/\sigma^2$ and $A\,[g(\mu_t)] = \mu^2/2\sigma^2$. One can verify that $\mu_t = \sigma^2 X_t$, so $A(X_t) = \sigma^2 X_t^2/2$, with first and second derivatives μ_t and σ^2, respectively.

2.1. Related models

One of the most frequently used specifications in the area of discrete-valued time series is the Generalized Autoregressive Moving Average model, GARMA [3]. Here, the distribution of the process is usually assumed to be the one-parameter exponential family (1). From equation (2), the GARMA model is obtained when

$k = 0$, by setting $g \equiv \bar{g} \equiv h$ and $\nu_t = 1$, so that,

$$g(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{p} \phi_j g(Y_{t-j}) + \sum_{j=1}^{q} \theta_j \left[g(Y_{t-j}) - g(\mu_{t-j}) \right] , \tag{5}$$

where $\boldsymbol{\alpha} = \left(1 - \sum_{j=1}^{p} \phi_j B^j\right) \boldsymbol{\beta}$, $\boldsymbol{\beta}$ is a vector of constants and B is the lag operator. By rearranging the constant in terms of $\boldsymbol{\beta}$ we obtain the equation (3) of [3].

A suitable extension of the GARMA model (5), the martingalized GARMA (M-GARMA), has recently been introduced by [44]; it is derived from (2) by setting $k = 0$, $g(\mu_t) \equiv \bar{g}(\mu_t) = \mathrm{E}[h(Y_t)\,|\mathcal{F}_{t-1}]$ and $\nu_t = 1$:

$$\bar{g}(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{p} \phi_j h(Y_{t-j}) + \sum_{j=1}^{q} \theta_j \left[h(Y_{t-j}) - \bar{g}(\mu_{t-j}) \right] . \tag{6}$$

The relevant feature of the model is that it allows the residuals ε_t to be a martingale difference sequence, i.e. $\mathrm{E}(\varepsilon_t | \mathcal{F}_{t-1}) = 0$.

Another similar model has been developed by [36], [34] and [9] with the name GLARMA model; here again the distribution is the exponential family (1). We can write the GLARMA model (2) by setting $p = 0$, h as the identity and $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)\,|\mathcal{F}_{t-1}] = \mathrm{E}(Y_t\,|\mathcal{F}_{t-1}) = \mu_t$:

$$g(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j g(\mu_{t-j}) + \sum_{j=1}^{\tilde{q}} \theta_j \varepsilon_{t-j} , \tag{7}$$

where $\boldsymbol{\alpha} = \left(1 - \sum_{j=1}^{k} \gamma_j B^j\right) \boldsymbol{\beta}$. Here $\tilde{q} = \max(k, q)$ and $\theta_j = \gamma_j + \tau_j$ for $j = 1, \ldots, \tilde{q}$, where τ_j are free parameters. The formulation of the constant term in equation (7) as a function of $\boldsymbol{\beta}$ is equivalent to equation (13) in [17], the alternative definition of the GLARMA model originally introduced in [9]. Note that here, if $\nu_t = \sigma_t$, then the prediction error $\varepsilon_t = (Y_t - \mu_t)/\nu_t$ is a white noise process with unit variance.

Another promising stream of literature is due to [20], who introduced Poisson autoregression, henceforth Pois AR, which is obtained when (1) is $Pois(\mu_t)$, with $f(X_t) = \log(X_t)$, and in equation (2), we have $q = 0$ and $g \equiv h$: *identity*:

$$\mu_t = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \mu_{t-j} + \sum_{j=1}^{p} \phi_j Y_{t-j} . \tag{8}$$

The parameters in equation (8) are constrained in the positive real line. A variant of (8) is the log-linear Poisson autoregression, henceforth Pois log-AR, [22] which is obtained by (2) when $q = 0$, $f(X_t) = X_t$, $g(\mu_t) = \log(\mu_t)$ and $h(Y_t) = \log(Y_t + 1)$

$$\log(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \log(\mu_{t-j}) + \sum_{j=1}^{p} \phi_j \log(Y_{t-j} + 1) . \tag{9}$$

For Poisson data, the GARMA model (5) with identity or log links corresponds to a constrained Poisson autoregression where $\gamma_j = -\theta_j$ and ϕ_j is replaced by $\phi_j + \theta_j$, in equations (8) or (9). A model like (9) could also be used for Negative Binomial data, by rewriting the distribution in terms of the expected value parameter μ_t [5]:

$$q\left(Y_t|\mathcal{F}_{t-1}\right) = \frac{\Gamma(\nu + Y_t)}{\Gamma(Y_t+1)\Gamma(\nu)} \left(\frac{\nu}{\nu+\mu_t}\right)^{v} \left(\frac{\mu_t}{\nu+\mu_t}\right)^{Y_t} \tag{10}$$

where ν is the dispersion parameter (if integer, it is also known as the number of failures) and the usual probability parameter would be $p_t = \frac{\nu}{\nu+\mu_t}$. The distribution (10) with model (9) is obtained from the distribution (1), by setting the non-canonical link $g(\mu_t) = \log(\mu_t)$ and $Q_t = \log(1-p_t)$, rewritten as $f(X_t) = X_t - \log(\nu + e^{X_t})$, with $A(X_t) = -\nu \log\left(\frac{\nu}{\nu+e^{X_t}}\right)$ and $d(Y_t) = \log\frac{\Gamma(\nu+Y_t)}{\Gamma(Y_t+1)\Gamma(\nu)}$.

The Binary Autoregressive Moving Average (BARMA) model ([27, 38]), introduced for Binomial data, is obtained when (1) is $Bin(a, \mu_t)$, where a is known and the probability parameter $p_t = \mu_t/a$, and, in (2), $\gamma = 0$, $h : identity$ ($\bar{g}(\mu_t)$ reduces to μ_t) and $c = 0$. Then

$$g(\mu_t) = \mathbf{Z}_t^T\boldsymbol{\alpha} + \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j \left[Y_{t-j} - \mu_{t-j}\right] . \tag{11}$$

Even though this model is designed for use with Binomial distributions, so typically g : *logit* or g : *probit*, in general, the link function g can be any suitable function.

Although the class of models (1)-(14) is quite general, all the link functions involved are required to be continuous and bijective. Then, the framework presented in the current contribution does not include, for example, the Poisson threshold model, as defined in [29, eq. 9] and [14, Sec. 1.2.1].

2.2. New model specifications

Other models of potential interest, which are not explicitly specified in the existent literature, are instead encompassed in the framework (1)-(2). We discuss a class of *g*link-ARMA models. As a relevant instance, consider the log-ARMA model

$$\begin{aligned}\log(\mu_t) = \mathbf{Z}_t^T\boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \log(\mu_{t-j}) + \sum_{j=1}^{p} \phi_j \log(Y_{t-j}+1) \\ + \sum_{j=1}^{q} \theta_j \left[\frac{\log(Y_{t-j}+1) - \bar{g}(\mu_{t-j})}{\nu_{t-j}}\right]\end{aligned} \tag{12}$$

where $f(X_t) = X_t$, $\bar{g}(\mu_t) = \mathrm{E}\left[\log(Y_t+1)|\mathcal{F}_{t-1}\right]$ and $\nu_t = \sqrt{\mathrm{V}\left[\log(Y_t+1)|\mathcal{F}_{t-1}\right]}$. The model (12) detects the autoregressive effect of the past lags of Y_t and also

accounts for a long past feedback effect, via lags of μ_t; then, a white noise prediction error $\varepsilon_t = [(\log(Y_t + 1) - \bar{g}(\mu_t))/\nu_t]$ is added to the functional transformation of the data, where $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{V}(\varepsilon_t) = 1$. The same model (12), when (1) is $Bin(a, \mu_t)$, is recovered by setting the non-canonical link $X_t = g(\mu_t) = \log(\mu_t)$ and $Q_t = \log\left(\frac{p_t}{1-p_t}\right) = \log\left(\frac{\mu_t}{a-\mu_t}\right)$, rewritten as $f(X_t) = X_t - \log(a - e^{X_t})$, with $A(X_t) = a \log\left(\frac{a}{a-e^{X_t}}\right)$ and $d(Y_t) = \log\binom{a}{Y_t}$. Along the same lines, a logit-ARMA model can be specified for Binomial data as a combination of the BARMA model from [27] and an autoregressive component:

$$\log\left(\frac{\mu_t}{a-\mu_t}\right) = \mathbf{Z}_t^T\boldsymbol{\alpha} + \sum_{j=1}^{k}\gamma_j \log\left(\frac{\mu_{t-j}}{a-\mu_{t-j}}\right) + \sum_{j=1}^{p}\phi_j Y_{t-j} + \sum_{j=1}^{q}\theta_j\left[\log(Y_{t-j}+1) - \bar{g}(\mu_{t-j})\right] \tag{13}$$

where, in equation (1) we have $f(X_t) = X_t$ where the canonical link is $X_t = g(\mu_t) = \log\left(\frac{\mu_t}{a-\mu_t}\right)$, with $A(X_t) = a\log(1+e^{X_t})$ and $d(Y_t) = \log\binom{a}{Y_t}$. A similar model can also be specified by replacing the *logit* function with the *probit* link function.

The usefulness of the specifications (12)-(13) can mainly be exploited when a closed form expression is available for the conditional expectation $\bar{g}(\mu_t)$ (and possibly for the standard deviation ν_t). For example, if the distribution of $Y_t|\mathcal{F}_{t-1}$ is Log-normal (μ_t, σ^2), then $\bar{g}(\mu_t) = \mathrm{E}\left[\log(Y_t+1)|\mathcal{F}_{t-1}\right] = \log(\mu_t) - 1/2\sigma^2$. For a comprehensive discussion on the closed form solutions see [44]. In the case of Binomial or Poisson data, though, such closed forms are not available and it seems reasonable to use an approximation based on the Taylor expansion around the mean μ_t, like $\bar{g}(\mu_t) = \mathrm{E}\left[h(Y_t)|\mathcal{F}_{t-1}\right] \approx h(\mu_t)$. This would reduce models (12)-(13) to an interesting reparametrized version of the log-AR model described in equation (9).

Despite the wide use of the Poisson model for count data and the default Negative Binomial alternative to account for over-dispersion, both choices fail when data present under-dispersion or an excess of zero value observations. For instance, the use of generalized Poisson distributions is quite popular, as well as the use of alternative link functions, when the canonical ones are not appropriate to the scientific ground; see [35] for a discussion. Choosing the link function in non-linear models is a relevant issue in case of over-dispersed and under-dispersed count data as explicitly debated in [30]. Generalized approaches to accommodate specific data structures may benefit from a flexible specification of *g*link-ARMA type models.

3. Stochastic properties

In the following, we shall focus our attention to the baseline model in the class, obtained by setting $k = p = q = 1$ in equation (2) with no covariates ($\mathbf{Z}_t^T\boldsymbol{\alpha} = \alpha$)

and unitary scaling sequence, $\nu_t = 1$ for $t \in T$:

$$g(\mu_t) = \alpha + \gamma\, g(\mu_{t-1}) + \phi\, h(Y^*_{t-1}) + \theta \left[h(Y^*_{t-1}) - \bar{g}(\mu_{t-1}) \right] , \qquad (14)$$

where the function Y^*_t modifies the values of Y_t to lie within the domain of $h(\cdot)$. Establishing stochastic properties and inferential results for model (2) is typically challenging beyond the order one; see, for example, [14, Sec. 4]. Indeed, in the vast majority of the contributions on observation-driven models for integer-valued data the theoretical results are derived for first order models. Remarkable exceptions are referred for the simpler linear INGARCH model [18]. For the same model, Poisson quasi maximum likelihood inference is discussed in [1]. In [29, pg. 813], the authors seem to show a way to extend the ergodicity results of the GARMA(1,1) model, to the general lag (p,q) order, but in fact, such an extension is possible after perturbing the original model with a stochastic noise, which is equivalent to making a correction for continuity on the starting integer-valued $\{Y_t\}$ process. For these reasons, the results of the present contribution are derived for the framework (14).

In Remark 1 we discuss an extension which includes non-unitary scaling sequences. In addition, although the inclusion of time-varying covariates, $\mathbf{Z}_t$, in model (14) may determine it to be, in general, non-stationary [29, pg. 810], the addition of a non-time-varying random vector, $\mathbf{Z}$, to (14) will keep all the results of the present paper unaffected. Note that in the first order observation-driven model (14) the series μ_t can be determined recursively by knowing only the starting point μ_0 and the observations $Y_0, \ldots, Y_{t-1}$. This major simplification is lost in the case of models with a lag order greater than the first.

3.1. Stationarity and ergodicity

The proof of the stability conditions is established for the baseline model (14) by showing the ergodicity of a first order Markov chain process. In the present section the random process $\{Y_t\}$ is not required to be distributed according to the exponential family (1) but to satisfy only relatively mild moment conditions, i.e. assumptions (A1)-(A2) below. Define $\mu_0 = \mu$, $g(\mu) = x$ and $\bar{g}(\mu) = \bar{g}(g^{-1}(x)) = \tilde{g}(x)$, where $\tilde{g}(\cdot) \equiv \bar{g} \circ g^{-1}(\cdot)$. In order to deal with different possible domains of the process $\{\mu_t\}$, we consider three separate cases:

Case 1: $q(Y_t|\mathcal{F}_{t-1})$ for $\mu \in \mathbb{R}$. The domain of g and h is $\mathbb{R}$ and $Y^*_t = Y_t$;
Case 2: $q(Y_t|\mathcal{F}_{t-1})$ for $\mu \in \mathbb{R}^+$ (or μ on one-sided open interval); the domain of g and h is $\mathbb{R}^+$ and $Y^*_t = \max\{Y_t, c\}$ for some $c \geq 0$;
Case 3: $q(Y_t|\mathcal{F}_{t-1})$ for $\mu \in (0, a)$ where $a > 0$ (or bounded open interval); the domain of g and h is $(0, a)$ and $Y^*_t = \min\{\max(Y_t, c), (a - c)\}$ for some $c \in [0, a/2)$.

Denote with $X = \{X_t\}_{t \in T}$ a Markov chain where $X_t = g(\mu_t)$ belongs to the state space S with σ-algebra $\mathcal{F}^X$ and define $P^t(x, A) = \mathrm{P}(X_t \in A \,|\, X_0 = x)$ for $A \in \mathcal{F}^X$ to be the t-step transition probability with initial state $X_0 = x$. Consider the following assumptions:

(A1) $\mathrm{E}(Y_t \mid \mu_t) = \mu_t$;
(A2) $\exists \delta > 0, r \in [0, 1+\delta)$ and $l_1, l_2 \geq 0$ s.t. $\mathrm{E}(|Y_t - \mu_t|^{2+\delta} \mid \mu_t) \leq l_1 |\mu_t|^r + l_2$;
(A3) g and h are bijective, increasing and
1. if $\bar{g}(\mu_t) = g(\mu_t)$,
   (a) $h : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $g : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $|\gamma| + |\phi| < 1$,
   (b) $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $(|\gamma| + |\phi|) \vee |\gamma - \theta| < 1$,
   (c) $h : (0, a) \mapsto \mathbb{R}$ and $g : (0, a) \mapsto \mathbb{R}$, $|\gamma - \theta| < 1$.
2. if $\bar{g}(\mu_t) \neq g(\mu_t)$ and $\tilde{g}(x)$ is Lipschitz with constant $L \leq 1$,
   (a) $h : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $g : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $|\gamma| + |\phi| < 1$,
   (b) $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $|\gamma| + (|\phi| \vee |\theta|) < 1$,
   (c) $h : (0, a) \mapsto \mathbb{R}$ and $g : (0, a) \mapsto \mathbb{R}$, $|\gamma| + |\theta| < 1$.

(A4) define $\pi_z(\cdot)$ as the distribution of $g(Y_t)$ conditional on $g(\mu_t) = z$; then, $\pi_z(\cdot)$ has the Lipschitz property $\sup_{w,z \in \mathbb{R}: w \neq z} \|\pi_w(\cdot) - \pi_z(\cdot)\|_{TV} / |w - z| < B < \infty$, where $\|\cdot\|_{TV}$ is the total variation norm.

Theorem 1 Suppose that $\{Y_t\}_{t \in T}$ has a distribution which satisfies (A1)-(A2), with the process $\{\mu_t\}_{t \in T}$ specified as in (14). Moreover, (A3)-(A4) hold. Then, $\{\mu_t\}_{t \in T}$ has a unique stationary distribution. This implies that $\{Y_t\}_{t \in T}$ is strict-sense stationary and ergodic.

The proof and some preliminary lemmata are postponed to the Appendix.

It can immediately be seen that in the special case where Y_t is distributed according to (1), Assumption (A1) automatically holds, because $\mu_t = \mathrm{E}(Y_t | \mathcal{F}_{t-1})$. For model (14), the σ-algebra generated by μ_t is a subset of $\mathcal{F}_{t-1}$, and for the properties of conditional expectations, $\mathrm{E}(Y_t | \mu_t) = \mathrm{E}[\mathrm{E}(Y_t | \mathcal{F}_{t-1}) | \mu_t] = \mu_t$. Assumption (A2) is a relatively mild moment condition generally satisfied for usual discrete distributions encompassed in (1), such as, e.g., Poisson and Binomial distributions, see also [29, Cor. 6-7]. Assumption (A3) involves several different conditions depending on three factors: (i) the specific model employed; (ii) the selected distribution, and (iii) the chosen link functions. To illustrate the implication of (A3) on existing models, we consider two representative examples.

Example 3 When a GARMA model (5) is applied (with $p = q = 1$ and no covariates) to Binomial data, we have $\bar{g}(\mu) = g(\mu)$, so that we fall in the case (A3).1. The domain of the observations involves (A3).1.(c), since $y \in (0, a)$ and $\mu \in (0, a)$, such that $g(\cdot) : (0, a) \mapsto \mathbb{R}$, where $g(y) = h(y)$. The monotonicity conditions on the shape of the link functions g and h in (A3) are quite standard. For instance, the logit link function $g(\mu) = \log(\mu/(a - \mu))$ is bijective and increasing. As $\gamma = 0$ in model (5), for the Binomial GARMA model, the stationarity condition $|\theta| < 1$ is obtained from (A3).

Example 4 For the GLARMA model (7) (with $k = q = \tilde{q} = 1$ and no covariates) and Poisson data with monotone increasing link function $g(\mu) =$

$\log(e^{\mu}-1)$, we have $h(y)=y$, i.e. $h(\cdot)$ is the identity, a monotone increasing function. Moreover, $\bar{g}(\mu_t)=\mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]=\mathrm{E}[Y_t|\mathcal{F}_{t-1}]=\mu_t$ is again the identity function, so $\bar{g}(\mu)\neq g(\mu)$ and $\tilde{g}$ is Lipschitz with constant $L\leq 1$, since $|\tilde{g}(x)-\tilde{g}(x^*)|=|\bar{g}(g^{-1}(x))-\bar{g}(g^{-1}(x^*))|=|g^{-1}(x)-g^{-1}(x^*)|\leq \max_{x\in\mathbb{R}}|\partial g^{-1}(x)/\partial x||x-x^*|\leq |x-x^*|$, where the last inequality follows by $\partial g^{-1}(x)/\partial x = e^x/(1+e^x)\leq 1$. This implies that the stationarity conditions for GLARMA are entailed by assumption (A3).2. In particular, in the Poisson case, (A3).2.(b) is involved, with $y\in\mathbb{R}^+$ and $\mu\in\mathbb{R}^+$. One observes that $g(\mu):\mathbb{R}^+\mapsto\mathbb{R}$ is concave and the same holds for $h(y)$. Since $\phi=0$ for model (7), we have that, for the Poisson GLARMA model, the stationarity condition $|\gamma|+|\theta|<1$ is obtained from (A3).

Many other results in the same spirit of Examples 3-4 can be obtained from Theorem 1, by selecting several different combinations of model, distribution, and link function, i.e. aspects (i)-(iii) previously discussed. Section 3.2 analyses the impact of the assumptions of Theorem 1 over the models introduced in Section 2.

As far as Assumption (A4) is concerned, though it might not be immediate to verify, it can usually be replaced, for the integer-valued distribution encompassed in (1), with an alternative condition, which is easier to check:

(A5) The distribution of the process $\{Y_t\}_{t\in T}$ is Poisson, Binomial, or Negative Binomial (with known number of trial/failure), and $g^{-1}(\cdot)$ is Lipschitz.

The equivalence of (A4) and (A5) has been proved in [29] for the Poisson and Binomial distribution; the proof for the Negative Binomial is reported in the Appendix. The required Lipschitz continuity of $g^{-1}(\cdot)$ is easily met for the usual link functions (e.g. logit, identity). However, there are exceptions, like the log link function. The modified log link function [29, eq 12], employed in Example 4, provides a viable alternative to avoid the problem.

Remark 1 Consider equation (14) with $\bar{g}(\mu_t)=\mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]$ and scaling sequence $\nu_t=\sigma(\mu_t)=\sqrt{\mathrm{V}[h(Y_t)\,|\mathcal{F}_{t-1}]}$, i.e.

$$g(\mu_t)=\alpha+\gamma\, g(\mu_{t-1})+\phi\, h(Y_{t-1})+\theta\varepsilon_t, \tag{15}$$

where ε_t, as in equation (3), is a white noise with unit variance. Under the conditions of the following corollary, the scaling sequence does not affect the stationarity conditions.

Corollary 1 Let $\nu_t=\sigma(\mu_t)$. Theorem 1 still holds true by replacing (14) with (15) if the function $\sigma(\cdot)$ is:

1. increasing for $\mu_t\in\mathbb{R}^+$ and decreasing for $\mu_t\in\mathbb{R}^-$;
2. increasing for $\mu_t\in\mathbb{R}^+$;
3. monotone with respect to μ_t;

depending on the domain of μ_t.

Corollary 1 follows from Theorem 1 in a way that is non-straightforward to prove; an outline of the proof is thus reported in the Appendix. Subsequent corollaries will come without proof. The conditions on ν_t are, in general, widely

satisfied. For example, if Y_t belongs to the exponential family in (4), $\sigma^2(\mu) = A''(X_t) = (g^{-1})'(g(\mu))$ where g is increasing by assumption, whereas $\sigma^2(\mu)$ is increasing since $(g^{-1})'$ is increasing; this holds as long as g is concave (g^{-1} is convex) which is true for $\mu > 0$. By contrast, $\sigma^2(\mu)$ is decreasing if $(g^{-1})'$ is decreasing which happens when g is convex: this is the case of $\mu < 0$, which is what was required.

Remark 2 It is worth noting that Assumption (A3) guarantees the existence of the r^{th} moment of all link functions $(h, g, \bar{g})$ provided that $\mathrm{E}|Y_t|^r < \infty$, for some integer r. Indeed, if $Y_t \in \mathbb{R}$, then $\mu_t \in \mathbb{R}$ (*Case 1*) and, since h is monotone increasing, concave in $\mathbb{R}^+$ and convex in $\mathbb{R}^-$, it is not hard to show that $|h(y)| \leq a_0 + a_1|y|$ for all $y \in \mathbb{R}$ and for some non-negative constants a_0, a_1. Analogous inequalities hold for generic integers $r > 1$, by the binomial theorem. Further details can be found in the Proof of Lemma 2 reported in the Appendix. The same arguments apply to g, since $|g(\mu)| \leq a_0 + a_1|\mu|$, for all $\mu \in \mathbb{R}$, which follows by $\mathrm{E}|\mu_t|^r \leq \mathrm{E}|Y_t|^r$. For the function $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]$, one has that $\mathrm{E}|\bar{g}(\mu_t)|^r \leq \mathrm{E}|h(Y_t)|^r$. When $Y_t \in \mathbb{R}^+$ (*Case 2*), the same conclusions follow. Finally, when $Y_t \in (0, a)$, for $a > 0$ (*Case 3*), the random variables are bounded.

3.2. Stochastic properties for relevant encompassed models

The results obtained in the previous section can be applied to specific models belonging to the unified framework (14) or (15), and in particular to the first order version of the novel models introduced in Section 2.2. We also specifically derive the stochastic properties of the related models discussed in Section 2.1, since for some of them the stochastic properties have not been fully addressed in the literature. Consider the one lag model (14).

As a proof of coherence in our findings, it is worth noting that, when $\gamma = 0$ and $g \equiv h \equiv \bar{g}$, Theorem 1 reduces to Theorem 5 in [29], providing results for the first order GARMA model

$$g(\mu_t) = \alpha + \phi\, g(Y^*_{t-1}) + \theta \left[g(Y^*_{t-1}) - g(\mu_{t-1})\right]. \tag{16}$$

Now we derive the stochastic properties for the first order BARMA model (11), such as

$$g(\mu_t) = \alpha + \phi Y_{t-1} + \theta \left(Y_{t-1} - \mu_{t-1}\right). \tag{17}$$

Corollary 2 Suppose that, conditional on $\mathcal{F}_{t-1}$, Y_t is $\mathrm{Bin}(n, \mu_t)$ with fixed number of trials n, link function $g : (0, a) \mapsto \mathbb{R}$ is bijective and increasing, g^{-1} is Lipschitz with constant not greater than 1 and $|\theta| < 1$. Then the process $\{\mu_t\}_{t\in T}$ defined in (17) has a unique stationary distribution. Hence, the process $\{Y_t\}_{t\in T}$ is strictly stationary and ergodic.

Note that for Binomial distribution (A1)-(A2) hold. Here, the conditions (A3) and (A5) on g and g^{-1} are clearly satisfied for usual link functions, like the logit.

To the best of our knowledge, no results are available for strict stationarity in the GLARMA model, apart from the simplest case when $k = 0$, $q = 1$ [9, 17]. Define the GLARMA(1,1) model as

$$g(\mu_t) = \alpha + \gamma\, g(\mu_{t-1}) + \theta\varepsilon_t. \tag{18}$$

Corollary 3 Suppose that $\{Y_t\}_{t\in T}$ is distributed according to (1). The process $\{\mu_t\}_{t\in T}$ in (18) has a unique stationary distribution and $\{Y_t\}_{t\in T}$ is strictly stationary and ergodic, if (A2) holds and

1. g is bijective and increasing, and
   - (a) $g : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $|\gamma| < 1$;
   - (b) $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $|\gamma| + |\theta| < 1$;
   - (c) $g : (0, a) \mapsto \mathbb{R}$, $|\gamma| + |\theta| < 1$.
2. g^{-1} is Lipschitz with constant not greater than 1.

In the GLARMA model, the conditional distribution of $\{Y_t\}_{t\in T}$ is part of the exponential family, then (A1) holds true. Instead, (A3) and (A5) reduce to conditions 1 and 2, which clearly are widely satisfied for the usual link functions. In practical applications, the condition on the coefficients of the model is required to establish its stationarity.

The proof of stationarity for the one lag M-GARMA model from (6) given in [44] only holds for continuous variables. We generalize the result by deriving the conditions for stationarity also for the case of discrete variables. They are shown to be equivalent to those available for the GARMA model. This is reasonable since the former is a special case of the latter. We now move to strict-stationarity and ergodicity results for some of the novel models presented in Section 2.2.

We discuss the result for model (8), with no covariates and $p = k = 1$. Note that the components of the model are all non-negative, $\bar{g}$ is not present, and $\theta = 0$. Indeed, conditions (A3)1.(b) and (A3)2.(b) in Theorem 1 coincide, providing the following conclusion.

Corollary 4 Suppose that, conditional on $\mathcal{F}_{t-1}$, Y_t is Pois(μ_t) and $\gamma + \phi < 1$. Then the process $\{\mu_t\}_{t\in T}$ defined in (8) has a unique stationary distribution. Hence, the process $\{Y_t\}_{t\in T}$ is strictly stationary and ergodic.

These results are not new in the literature [20, 16].

Corollary 5 Suppose that $\{Y_t\}_{t\in T}$ is distributed as in (1), $\tilde{g}(x)$ is Lipschitz with constant $L \leq 1$, (A2), (A4) hold and $|\gamma| + (|\phi| \vee |\theta|) < 1$. Then the process $\{\mu_t\}_{t\in T}$ defined in (12), with $k = p = q = 1$, has a unique stationary distribution. Hence, the process $\{Y_t\}_{t\in T}$ is strictly stationary and ergodic.

Assumption (A1) is met for the distribution (1). The condition (A3) on the shape of the link function holds here, as $g(\mu) = \log(\mu)$. However, the Lipschitz continuity on $\tilde{g}(\cdot)$ and the condition (A4) are required since $g^{-1}(\cdot)$ does not satisfy (A5). Note that the stationarity and ergodicity for the model (9), with no covariates and $p = k = 1$, are established as a special case of Corollary 5, with $q = \theta = 0$.

Corollary 6 Suppose that $\{Y_t\}_{t\in T}$ is distributed as $\text{Bin}(n, \mu_t)$ with fixed number of trials n, $\tilde{g}(x)$ is Lipschitz with constant $L \leq 1$ and $|\gamma| + |\theta| < 1$. Then the process $\{\mu_t\}_{t\in T}$ defined in (13), with $k = p = q = 1$, has a unique stationary distribution. Hence, the process $\{Y_t\}_{t\in T}$ is strictly stationary and ergodic.

For Binomial distributions (A1)-(A2) hold and the conditions (A3) and (A5) are satisfied for the logit link function. For space constraints, we do not show other examples. However, based on the theoretical results developed for this flexible framework, stationarity and ergodicity can be directly established for a wide class of models under several discrete distributions.

For a better readability we have summarized, in Table 1, the stationarity and ergodicity conditions found in this section concerning parameters of already existing models, so as to allow the current results obtained by Theorem 1 to be compared with the ones previously available in the literature. The last column reports the papers where the previous conditions were established. For the log-AR model the conditions determined by [14] are different and less restrictive than ours. This is because the authors employed a different approach. However, when the coefficients are positive the two conditions are equivalent. All the other results found in the current contribution are new or equal the ones already existing in the literature.

TABLE 1
Stationarity and ergodicity conditions for the parameters of models presented in Section 2.1.

Models	Condition	Count	Binomial	Work
GARMA	Previous	$\lvert\phi\rvert \vee \lvert\theta\rvert < 1$	$\lvert\theta\rvert < 1$	[29]
	Current	$\lvert\phi\rvert \vee \lvert\theta\rvert < 1$	$\lvert\theta\rvert < 1$	
M-GARMA	Previous	–	–	
	Current	$\lvert\phi\rvert \vee \lvert\theta\rvert < 1$	$\lvert\theta\rvert < 1$	
BARMA	Previous		$\lvert\theta\rvert < 1^*$	[27]
	Current		$\lvert\theta\rvert < 1$	
GLARMA	Previous	–	–	
	Current	$\lvert\gamma\rvert + \lvert\theta\rvert < 1$	$\lvert\gamma\rvert + \lvert\theta\rvert < 1$	
AR	Previous	$\gamma + \phi < 1$		[16]
	Current	$\gamma + \phi < 1$		
log-AR	Previous	$\lvert\gamma + \phi\rvert \vee \lvert\gamma\rvert \vee \lvert\phi\rvert < 1$		[14]
	Current	$\lvert\gamma\rvert + \lvert\phi\rvert < 1$		

Note: – not available. * without proof.

4. Quasi-maximum likelihood inference

The aim of this section is to establish the asymptotic theory of the quasi maximum likelihood estimator of the parameter $\rho = (\alpha, \gamma, \phi, \theta)$. More precisely we develop asymptotic results in the three following cases: (*i*) misspecified MLE: misspecification occurs in the distribution (1) and/or in the model (2), and (*ii*) QMLE: misspecification occurs in the distribution (1), (*iii*) correctly specified MLE. Specifically, strong consistency is derived in the three cases; asymptotic normality is derived for the QMLE and the MLE. Finite sample properties are explored through an extensive simulation study, as well as the performance of

information criteria for model selection. Tables including detailed and numerical results are postponed to the Appendix.

4.1. Asymptotic properties

Assume that the variables in the process $\{Y_n\}_{n\in\mathbb{Z}}$ are integer-valued. Let (Λ, d) be a compact metric set of parameters, with suitable metric $d(\cdot)$, and consider $\tilde{\alpha}, \tilde{\delta} \in \mathbb{R}^+$. Then, $\Lambda = \left\{\rho = (\alpha, \gamma, \phi, \theta) \in \mathbb{R}^4 : |\alpha| \leq \tilde{\alpha}, |\delta| = |\phi + \theta| \leq \tilde{\delta}\right\}$ is the parameter set. We make explicit the dependence of the conditional distribution (1) from the mean process by using the notation $q(y_t|\mathcal{F}_{t-1}) = q(X_t; y_t)$. Let $g^\rho\langle Y_{-\infty:t}\rangle$ be a stationary ergodic random process, not necessarily equal to the process $X_t = g(\mu_t)$ in (14),

$$g^\rho\langle Y_{-\infty:t}\rangle = \alpha + \gamma g^\rho\langle Y_{-\infty:t-1}\rangle + \phi h(Y_{t-1}) + \theta[h(Y_{t-1}) - \tilde{g}(g^\rho\langle Y_{-\infty:t-1}\rangle)]\,, \quad (19)$$

and its sample counterpart is denoted by $g^\rho\langle y_{1:t-1}\rangle(x)$, where x is the starting value of the chain $g^\rho\langle\cdot\rangle$. The notation $g^\rho\langle y_{s:t}\rangle(x) = g^\rho_{y_t} \circ g^\rho_{y_{t-1}} \circ \cdots \circ g^\rho_{y_s}(x)$, $s \leq t$ is the so-called Iterated Random Function (IRF), see [13], with

$$g^\rho_{y_1}(x) = \alpha + \gamma x + \phi h(y_0) + \theta[h(y_0) - \tilde{g}(x)]\,. \quad (20)$$

It is worth noting that in the special case of a correctly specified model, $X_0 = g^\rho\langle Y_{-\infty:0}\rangle$ and equation (19) reduces exactly to the process in equation (14). Let us define the log-likelihood function as follows:

$$L^\rho_{n,x}\langle Y_{1:n}\rangle \coloneqq n^{-1} \log\left(\prod_{t=1}^{n} q(g^\rho\langle Y_{1:t-1}\rangle(x); Y_t)\right),$$

whose associated maximum likelihood estimator is

$$\hat{\rho}_{n,x} = \arg\max_{\rho\in\Lambda} L^\rho_{n,x}\langle Y_{1:n}\rangle\,. \quad (21)$$

We specify the following assumptions:

(H1) $\mathrm{E}[\log|A'(g^\rho\langle Y_{-\infty:0}\rangle)|]_+ < \infty$, $\mathrm{E}[\log|f'(g^\rho\langle Y_{-\infty:0}\rangle)|]_+ < \infty$, $\mathrm{E}|Y_0| < \infty$,
(H2) $\mathrm{E}[A'(g^\rho\langle Y_{-\infty:0}\rangle)^4] < \infty$, $\mathrm{E}[f'(g^\rho\langle Y_{-\infty:0}\rangle)^4] < \infty$,
$\mathrm{E}[A''(g^\rho\langle Y_{-\infty:0}\rangle)^4] < \infty$, $\mathrm{E}[f''(g^\rho\langle Y_{-\infty:0}\rangle)^4] < \infty$, $\mathrm{E}(Y_0^4) < \infty$,

which are mild conditions for the existence of moments, in general immediate to verify; see the related section A.8 in the Appendix for some relevant examples.

Firstly, consistency for the misspecified MLE is proven, then the other two ML estimators are derived as special cases of it.

Theorem 2 Suppose Theorem 1 and (H1) hold. Then, $\lim_{n\to\infty} d(\hat{\rho}_{n,x}, \mathrm{P}_\star) = 0$, a.s., $\forall x \in S$, with $\mathrm{P}_\star \coloneqq \arg\max_{\rho\in\Lambda} \mathrm{E}\{Y_0\, f[g^\rho\langle Y_{-\infty:0}\rangle] - A[g^\rho\langle Y_{-\infty:0}\rangle] + d(Y_0)\}$.

Here, the almost sure limit is taken under the stationary distribution of $\{Y_t\}_{t\in T}$. The proof is in the Appendix. Now the special case of correctly specified MLE is treated. Let us denote Λ^0 as the interior of the set Λ.

Theorem 3 Assume that $\{Y_n\}_{n\in\mathbb{Z}}$ is distributed according to (1) and satisfies the recursion (14), with parameters $\rho_\star \in \Lambda^0$. Moreover, assume that Theorem 2 holds. Then, for all $x \in S$, $\lim_{n\to\infty} \hat{\rho}_{n,x} = \rho_\star$, a.s.

We need to show that $\mathrm{P}_\star = \{\rho_\star\}$. The proof is postponed to the Appendix. The asymptotic consistency of QMLE is now established.

Corollary 7 Assume that $\{Y_n\}_{n\in\mathbb{Z}}$ satisfies the recursion (14), with parameters $\rho_\star \in \Lambda^0$ and $\mu = A'(x_\star)$. Moreover, assume that Theorem 2 holds. Then, for all $x \in S$,

$$\lim_{n\to\infty} \hat{\rho}_{n,x} = \rho_\star, \quad \text{a.s.} \tag{22}$$

where $\{x_\star\}$ is the maximum of the function $\int P(x_\star, dy) \log q(x, y)$.

In practice, $\mu = A'(x_\star)$ states that the mean function has to be correctly specified regardless the true data generating process. The proof is analogous to Theorem 3 and follows directly by Theorem 4.1 and [15, Thr 4.1]. Finally, we investigate the conditions under which the QMLE (22) is asymptotically normally distributed for the model (14).

Theorem 4 Assume that Corollary 7 and (H2) hold. Moreover, assume that $\mathcal{J}(\rho_\star)$ is non-singular. Then, $\sqrt{n}(\hat{\rho}_{n,x} - \rho_\star) \xRightarrow{D} \mathrm{N}(0, \mathcal{J}(\rho_\star)^{-1}\mathcal{I}(\rho_\star)\mathcal{J}(\rho_\star)^{-1})$, where

$$\mathcal{I}(\rho_\star) \coloneqq \mathrm{E}\left[\left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right) \left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right)' \left(\frac{\partial}{\partial x} \log q\left(g^{\rho_\star}\langle Y_{-\infty:0}\rangle, Y_1\right)\right)^2 \right],$$

$$\mathcal{J}(\rho_\star) \coloneqq \mathrm{E}\left[\left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right) \left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right)' \frac{\partial^2}{\partial x^2} \log q\left(g^{\rho_\star}\langle Y_{-\infty:0}\rangle, Y_1\right)\right].$$

The proof relies on the argument of [15, Thr 4.2] and follows the fashion and the notation used in the proof of Theorem 2, thus it is postponed to the Appendix. It goes without saying that for correctly specified MLE, equation (21) is the exact MLE and $\mathcal{J}(\rho_\star) = \mathcal{I}(\rho_\star)$ in Theorem 4, providing the standard ML inference.

4.2. Finite sample properties

Finite sample properties of MLE and QMLE are explored through a simulation study which considers some models illustrated in Section 2.1. Tables including the details of the numerical results are stored in Section B of the Appendix. All the results are based on $s = 1,000$ replications, with different configurations of the parameters and increasing sample size $n = (200, 500, 1,000)$. A correctly specified MLE has been estimated on data coming from Bernoulli or Poisson distributions, across several models. For QMLE, data are generated from a Geometric distribution, with Poisson distribution fitted instead, for GARMA and log-AR models. For Poisson and Geometric data, the log-link is employed $g(\mu_t) = \log(\mu_t)$; instead, for the Bernoulli one, the logit $g(\mu_t) = \log(\mu_t)/\log(1-\mu_t)$ is specified. For all the models involved, the mean of the estimators approaches

the true value, for both the well-specified MLE and QMLE. Some convergence problems arise for the BARMA model, but the standard error and the bias still tend to reduce by increasing n; this gives evidence of convergence, although at a slower rate. Turning to asymptotic normality, evidence of normality emerges from the Kolmogorov-Smirnov (KS) test, even when the sample size is small. The outcomes are in line with those of [15]. These results are coherent with the theory presented so far.

In summary, Table A-1 in the Appendix reports the estimation results for the GLARMA model when the data come from a Bernoulli distribution. The estimates tend to be closer to the true value of the parameters as the sample size increases, which confirms the consistency of the estimators. Consequently, the bias is also reduced. Moreover, the estimates are significant at the usual levels and the true value of the parameters falls into the confidence intervals. The KS tests do not reject the normality of the estimators even with a small sample size. The same comments hold true for all the combinations of parameters employed. Similar results are obtained in Tables A-2 and A-3 which show the outcome of simulations for the GARMA and log-AR models, respectively, performed on data generated from the Geometric distribution in (10), but with Poisson distribution fitted instead (QMLE). The GARMA model seems to be more accurate on the approximation of the true values but some problems with the KS test are found when a non-stationary region for the parameters $\rho = (0.5, 0.4, 1.2)$ is investigated. Instead, the log-AR model could not be estimated in non-stationary regions of the parameters.

4.3. Model selection

A crucial aspect in empirical applications is model selection. In likelihood inference, model selection is typically carried out based on information criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC). To assess the effectiveness of AIC and BIC for selecting the most appropriate model for the data at hand, we carry out an extensive simulation study with competing one lag models log-AR, GARMA, and GLARMA for Poisson data. The last two are also computed, together with the BARMA model, for Binomial data. In extreme synthesis, when the sample size n is small, the selection for some models can perform poorly, but when n is sufficiently large ($n \geq 500$), all the models allow the selection of the right data generating model with high probability.

We simulate the first order log-AR, GARMA, and GLARMA models for $Y_t|\mathcal{F}_{t-1}$ distributed according to a $Pois(\mu_t)$, with $(\alpha, \phi, \theta, \gamma) = (0.2, 0.4, 0.2, 0.3)$, number of repetitions $S = 1,000$ and sample sizes $n = (250, 500, 1,000)$. The same is done by generating data from the first order BARMA, GARMA and GLARMA models, with $Bin(5, p_t)$, $p_t = \mu_t/a$ and $g(\mu_t) = \log(\mu_t)/\log(a - \mu_t)$. For the GARMA model, $g(y_t^\star) = \log(y_t^\star)/\log(1-y_t^\star)$, $y_t^\star = \min(\max(y_t, c), 5-c)$ and $c = 0.1$, whereas, in the GLARMA model, $s_t = \sqrt{5p_t(1-p_t)}$. For each distribution, we generate S times a vector of data with length n from one model,

TABLE 2
Frequency (%) of correct selection for AIC.

	Binomial			Poisson		
n	BARMA	GARMA	GLARMA	log-AR	GARMA	GLARMA
200	62.3	97.2	60.0	53.6	99.2	95.1
500	74.4	99.7	58.0	70.5	99.9	99.4
1000	83.8	100	81.0	85.6	100	100

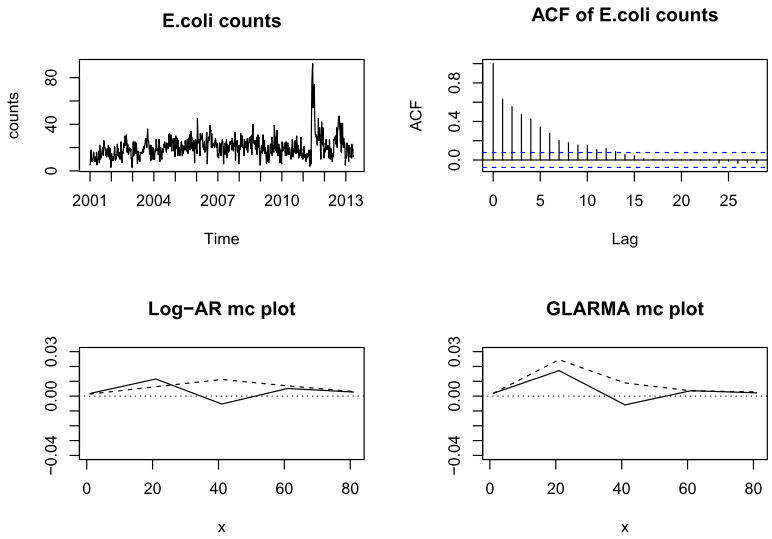

FIG 1. *Top-left: Escherichia coli counts. Top-right: ACF. Bottom-left: mc plot for log-AR model. Bottom-right: mc plot for GLARMA model. Dashed line is Poisson. Black line is NB.*

then the data generated are employed in the estimation of all the three models. The Akaike and the Bayesian information criteria are computed for each model. Finally, the frequency of correct selection over the S repetitions is established, counting the percentage of the number of times the information criteria selected the model truly employed to generate the data. The same procedure is replicated for all the models. The results for the AIC are summarized in Table 2 (results for the BIC are identical).

For the Poisson distribution, the results are excellent in the GARMA and the GLARMA models. The log-AR seems to show a slower convergence towards the right model, but it reaches a satisfactory result with increasing n. The same holds, in the case of Binomial data, for the BARMA and GLARMA models. Finally, the GARMA model also works very well for the Binomial distribution.

5. Application on disease cases of Escherichia coli in North Rhine-Westphalia

We consider a test-bed time series, i.e. the weekly number of reported disease cases caused by Escherichia coli in the state of North Rhine-Westphalia (Germany) from January 2001 to May 2013. The data can be found in the `R` package `tscount` [28]. The time series has a time length $n = 646$ and is plotted in Figure 1, with its sample autocorrelation function (ACF). There is a temporal correlation which spreads over several lags with a greater magnitude compared to the data set in the previous example. The slow decay of the ACF suggests the use of a feedback mechanism.

For the data generating process we assume both the Poisson and the Negative Binomial (NB) distribution in equation (10), where $\nu > 0$ is the dispersion parameter and μ_t is the conditional expectation. Indeed, equation (10) is defined in terms of mean rather than of the probability parameter $p_t = \frac{\nu}{\nu+\mu_t}$ and it accounts for overdispersion in the data as $\mathrm{V}(Y_t|\mathcal{F}_{t-1}) = \mu_t\,(1 + \mu_t/\nu) \geq \mu_t$. We fit the following models

$$\begin{aligned}
&\text{log-AR:} && \log(\mu_t) = \alpha + \phi \log(y_{t-1} + 1) + \gamma \log(\mu_{t-1})\,,\\
&\text{GARMA:} && \log(\mu_t) = \alpha + \phi \log(y^\star_{t-1}) + \theta \left[\log(y^\star_{t-1}) - \log(\mu_{t-1})\right],\\
&\text{GLARMA:} && \log(\mu_t) = \alpha + \gamma \log(\mu_{t-1}) + \theta\varepsilon_t\,,
\end{aligned}$$

where $y^\star_t = \max\{y_t, c\}$ with $c = 0.1$, and $\varepsilon_t = (y_{t-1} - \mu_{t-1})s_{t-1}$. Different values of $0 < c < 1$ do not affect the estimates; while s_t is the square root of the conditional variance $s_t = \sqrt{\mu_t}$ for the Poisson distribution and $s_t = \sqrt{\mu_t\,(1 + \mu_t/\nu)}$ for the NB. In this likelihood-based framework, model selection is based on information criteria, such as AIC and BIC. The Quasi Information Criterion (QIC) introduced by [32] is also employed. It is a generalization of the AIC which takes into account the usage of a working quasi-likelihood instead of the true likelihood. QIC coincides with AIC in the case of well-specified models. QMLE has been carried out. The log-likelihood function of the Poisson and NB distributions is maximized by using a standard optimizer in `R` based on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. The score functions written in terms of predictor $x_t = \log \mu_t$ are:

$$\begin{aligned}
\chi_n(\rho) &= \frac{1}{n}\sum_{t=1}^{n}\Big(y_t - \exp x_t(\rho)\Big)\frac{\partial x_t(\rho)}{\partial \rho}\\
\chi_n(\rho) &= \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{(y_t+\nu)\exp x_t(\rho)}{\exp x_t(\rho) + \nu}\right)\frac{\partial x_t(\rho)}{\partial \rho}\,.
\end{aligned}$$

The solution of non-linear equation system $\chi_n(\rho) = 0$, if it exists, provides the QMLE of ρ (denoted by $\hat\rho$). In NB models, estimation of ν is also required. The moment estimator proposed in [6] is used:

$$\hat\nu = \left(\frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \hat\mu_t)^2 - \hat\mu_t}{\hat\mu_t^2}\right)^{-1} \tag{23}$$

where $\hat{\mu}_t = \mu_t(\hat{\rho})$ is the estimator from the Poisson model. Then, with $\nu = \hat{\nu}$ we estimate the NB model and obtain the new estimates for $\hat{\mu}_t$, plug them into (23), obtain a new value for $\hat{\nu}$, and repeat the procedure until a certain tolerance value is reached. The standard errors are computed from the "sandwich" estimators in Theorem 4; each quantity has been replaced by its sample counterpart. The results of the analysis are summarized in Table 3. For Log-AR, GARMA and GLARMA the whole set of parameters is significant at the 5% levels. The parameter $\hat{\nu}$ is generally around 10. All the information criteria select the NB GLARMA model as the best, in a goodness-of-fit sense. We then assess the adequacy of the fit.

TABLE 3
MLE results for Escherichia coli infection.

Models	$\hat{\alpha}$	$\hat{\phi}$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\nu}$	*AIC*	*BIC*	*QIC*
Pois log-AR	0.441	0.437	0.416	-	-	13.115	26.527	27.043
	(0.087)	(0.062)	(0.078)	-				
Pois GARMA	0.535	0.829	-	-0.418	-	13.134	26.546	27.371
	(0.095)	(0.031)	-	(0.079)				
Pois GLARMA	0.445	-	0.851	0.085	-	12.954	26.366	26.639
	(0.098)	-	(0.033)	(0.013)				
NB log-AR	0.546	0.400	0.419	-	10.030	12.633	26.045	12.432
	(0.102)	(0.05)	(0.073)	-				
NB GARMA	0.640	0.794	-	-0.420	9.865	12.641	26.053	12.576
	(0.111)	(0.036)	-	(0.074)				
NB GLARMA	0.483	-	0.839	0.142	10.892	**12.578**	**25.990**	**12.114**
	(0.110)	-	(0.036)	(0.019)				

The adequacy of the fit has been checked through the behaviour of the standardized Pearson residuals $e_t = [Y_t - \mathrm{E}(Y_t|\mathcal{F}_{t-1})] / \sqrt{\mathrm{V}(Y_t|\mathcal{F}_{t-1})}$, which is done by taking the empirical version $\hat{e}_t$ from the estimated quantities. If the model is correctly specified, the residuals should be white noise sequences with constant variance. This can be seen from the ACF, which in our case appears uncorrelated. [8] introduced a non-randomized version of Probability Integral Transform (PIT) for discrete data. It can be built based on the conditional cumulative distribution function

$$F(u|y_t) = \begin{cases} 0, & u \le P_t(y_t - 1) \\ \frac{u - P_t(y_t-1)}{P_t(y_t) - P_t(y_t-1)}, & P_t(y_t) \le u \le P_t(y_t - 1) \\ 1, & u \ge P_t(y_t) \end{cases} \tag{24}$$

where $P_t(\cdot)$ is the cumulative distribution function (CDF) at time t (in our case Poisson or NB). If the model is correct, $u \sim Uniform(0,1)$ and the PIT (24) will appear to be the cumulative distribution function of a $Uniform(0,1)$. The PIT (24) is computed for each realization of the time series y_t, $t = 1 \ldots, n$ and for values $u = j/J, j = 1, \ldots, J$, where J is the number of bins (usually equal to 10 or 20); then its mean $\bar{F}(j/J) = 1/n \sum_{t=1}^{n} F(j/J|y_t)$ is taken. The outcomes are probability mass functions, obtained in terms of differences $\bar{F}(\frac{j}{J}) - \bar{F}(\frac{j-1}{J})$; Figure 2 is a representative plot. The NB seems to be more appropriate for

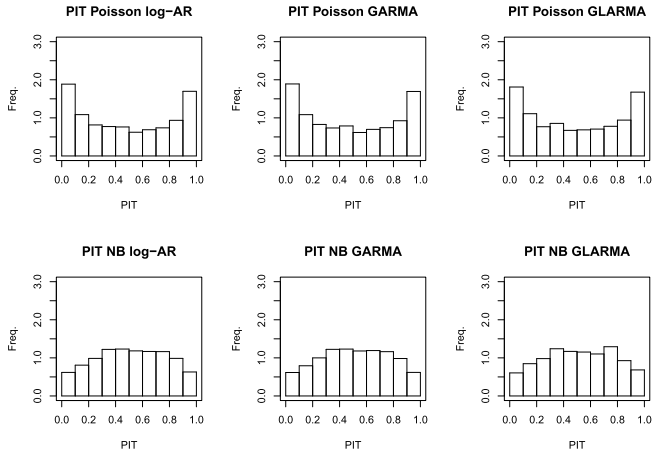

FIG 2. *PIT's for Escherichia coli counts. Top: Poisson. Bottom: NB.*

the data as its PIT's are quite near to $Uniform(0,1)$. Another control can be performed by using the probability and marginal calibration (mc), as defined in [24]. It compares the average of CDF selected, $\bar{P}(x) = 1/n \sum_{t=1}^{n} P_t(x)$, against the average of the empirical CDF, $\bar{G}(x) = 1/n \sum_{t=1}^{n} \mathbf{1}(y_t \leq x)$. The mc is plotted in Figure 1 for the log-AR and GLARMA models. In the other models the results are similar. Both distributions seem to show a good concordance with empirical distribution but the NB appears to perform better than the Poisson, especially for the larger quantiles.

In order to assess the predictive power, we refer to the concept of sharpness of the predictive distribution defined in [24]. It can be measured by some average quantities related to the predictive distribution, which takes the form $1/n \sum_{t=1}^{n} d(P_t(y_t))$, and $d(\cdot)$ is a scoring rule. We adopt the usual scoring rules employed in the literature: the logarithmic score (logs) $-\log p_t(y_t)$, where $p_t(\cdot)$ is the probability mass at the time t; the quadratic score (qs) $-2p_t(y_t) + \|p\|^2$, where $\|p\|^2 = \sum_{k=0}^{\infty} p_t^2(k)$; the spherical score (sphs) $-p_t(y_t)/\|p\|$ and the ranked probability score (rps) $\sum_{k=0}^{\infty}[P_t(k) - \mathbf{1}(y_t \leq k)]$. Numerical results for each model are collected in Table 4. The NB GLARMA model provides the best predictive performance for all the scoring rules analysed, and it is ultimately chosen, since it has been also selected by the information criteria.

6. Discussion

We developed statistical inference for a first order class of models for discrete time series which encompasses known models as well as new models of potential interest for the analysis of integer-valued time series. Stability conditions have been derived for the models in the class and a large family of probability distributions satisfying mild moment conditions. Consistency and asymptotic

TABLE 4
Predictive performance for Escherichia coli infection.

Models	Distribution	logs	qs	sphs	rps
log-AR	Poisson	3.5662	-0.0408	-0.2073	3.8480
	NB	3.3245	-0.0442	-0.2110	3.7960
GARMA	Poisson	3.5759	-0.0406	-0.2071	3.8591
	NB	3.3286	-0.0440	-0.2107	3.8105
GLARMA	Poisson	3.4859	-0.0420	-0.2097	3.7347
	NB	**3.2971**	**-0.0449**	**-0.2127**	**3.6801**

normality of the quasi maximum likelihood estimators have been also established, with the focus on the exponential family. The results about stochastic and inferential properties make any model belonging to the class fully applicable in practice.

An interesting extension of the present study may concern suitable specifications for multinomial data. Indeed, equation (1) describes an exponential family dealing only with one time-varying parameter μ_t. However, there are cases where several dynamic parameters characterize the distribution. Concerning discrete-valued data, this is the case of categorical random variables. An extension of the general framework to several parameters in the context of exponential families may be considered following the results on multinomial logistic models in [23], where exogenous covariates are also considered.

Another extension of potential interest may be the specification of multivariate discrete models. Recently, [21] established the multivariate discrete-valued extension of the (mixed) Poisson autoregression models (8)-(9). In line with such developments, the extension to a multivariate setting may represent a challenging research advance, though complications may arise at the modelling and at the inferential stage. For instance, as far as the stochastic properties are concerned, the coupling condition employed in this paper to show stationarity and ergodicity of the model (Lemma 4 in the Appendix) does not apply to multivariate processes. A possible direction to solve the problem may be based on the perturbation approach, as described by [21, Sec. 3.1-3.2]. In addition, the choice of a suitable multivariate version of the discrete probability mass function is non-trivial. Although several alternatives have been proposed in the literature, see the recent review in [19, Sec. 2], the choice of a suitable multivariate version of the discrete probability mass function remains a challenging problem. As a matter of fact, multivariate discrete probability mass functions have a complicated closed form and the associated likelihood inference is both theoretically and computationally cumbersome. Furthermore, in many cases, multivariate probability mass functions imply restricted models, of limited use in applications: see the discussion in [26] and [10]. A viable alternative may be the specification of joint distribution of the integer vector $\{Y_t\}$ by a copula approach, as described in [21, Sec. 2].

Along the lines traced in this discussion, we expect the specification of the broad class of models will provide useful enhancements to study the dynamic

trend of count and binary data.

Appendix A: Main proofs

A.1. Preliminary Lemmata for the Proof of Theorem 1

The proof of Theorem 1 is based on the following preliminary lemmata, stated with the same notation as the theorem. First, a small set (Lemma 1) and a drift condition are proved on the Markov chain $X_t = g(\mu_t)$ (Lemma 2); after that, the weak Feller property is established for the chain (Lemma 3), which proves the existence of a stationary distribution for $\{X_t\}_{t\in T}$. Then, the asymptotic strong Feller condition is verified (Lemma 4). Finally, the existence of a reachable point is shown (Lemma 5) and, by combining all these results, the uniqueness of the stationary distribution of the chain is proven. For definitions and properties invoked in the lemmata, see [2, pg. 54-55].

Let $\mathrm{E}_x(\cdot)$ denote the expectation under the probability $\mathrm{P}_x(\cdot)$ induced on the path space of the chain $\{X_t\}_{t\in T}$ when the initial state X_0 is deterministically equal to x. Consider the following drift condition $\forall x \in S$:

$$\mathrm{E}_x V(X_1) \leq \eta V(x) + b\mathbf{1}_{\{x\in A\}} \tag{A-1}$$

where $\eta \in (0,1)$, $b > 0$, $V : S \to [1,\infty)$ and $A \subset S$. We first prove that it is possible to select a set A which is small.

Lemma 1 Define a set $A = [-M, M] \subset S$, for some constant $M > 0$. Under assumptions (A1)-(A3), for the chain $\{X_t\}_{t\in T}$, A is small.

Proof Note that for any $x \in A$, $\mathrm{P}_x(Y_0 \in [a_1(M), a_2(M)]) > 3/4$ where

$$a_1(M) = g^{-1}(-M) - [4(l_1 \max\{|g^{-1}(-M)|, |g^{-1}(M)|\}^r + l_2)]^{1/(2+\delta)},$$
$$a_2(M) = g^{-1}(M) - [4(l_1 \max\{|g^{-1}(-M)|, |g^{-1}(M)|\}^r + l_2)]^{1/(2+\delta)}.$$

Given $X_0 = x$ and $\mu_0 = \mu = g^{-1}(x)$, we can write $\bar{g}(\mu) = \bar{g}(g^{-1}(x)) = (\bar{g} \circ g^{-1})(x) = \tilde{g}(x)$ where the composite function $\tilde{g}$ is still monotonic (and invertible), as a composition of monotonic functions. Then, with probability at least 3/4, $X_1 \geq \min\{b(a_1(M)), b(a_2(M))\} - |\gamma|M - |\theta||\tilde{g}(M)|$ and $X_1 \leq \min\{b(a_1(M)), b(a_2(M))\} + |\gamma|M + |\theta||\tilde{g}(M)|$, where $b(a) = \alpha + (\phi + \theta)h(a^*)$ and a^* is the operator * applied to a. This shows that A is a small set, by (A1)-(A3), in the fashion of [29, p. 812]. We omit the details. □

Lemma 2 Under assumptions (A1)-(A3), the chain $\{X_t\}_{t\in T}$ satisfies the drift condition (A-1).

Proof Consider the small set A in Lemma 1. We shall consider two states, $x > M$ and $x < -M$, each one, in turn, facing three different cases, according to the domain of the mean parameter μ_t, as described in Section 3.1. Set A as defined in Lemma 1. Take $V(x) = |x|$. We only prove the case where $x > M$ and the mean process $\mu_t \in \mathbb{R}$ (*Case 1*), as the other cases can be dealt with

in a similar manner. The interested reader can find full detailed proofs in [2, pg. 55-59].

We assume, without loss of generality, that $h(0) = 0$, since replacing $h(y)$ with $h(y) - h(0)$ simply changes the value of α. In this case, we assume that h is concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, so that there are constants $a_0, a_1 \geq 0$ such that $|h(y)| \leq a_0 + a_1|y|$ for all y; the same assumptions hold for g. Consider

$$\mathrm{E}_x V(X_1) \leq |\alpha| + |\gamma|\mathrm{E}_x|x| + |\phi|\mathrm{E}_x|h(Y_0)| + |\theta|\mathrm{E}_x|h(Y_0) - \bar{g}(\mu)|. \tag{A-2}$$

From equation (A-2), we need to show that

$$\mathrm{E}_x|h(Y_0)| \leq x + C\,. \tag{A-3}$$

When $h(\mu) \leq g(\mu)$, this holds from a result in [29, Sec. A.7,] by replacing $g(\cdot)$ by $h(\cdot)$. Instead, when $h(\mu) > g(\mu)$, the result is unchanged by applying the following inequality $h(\mu) = g(\mu+\delta) \leq g(\mu)+g(\delta)$, where $\delta > 0$, for the concavity of the functions involved in the same domain. Next, we show that the term $\mathrm{E}_x|h(Y_0) - \bar{g}(\mu)|$ in (A-2) is "small" relative to the linear term x. Specifically, we prove that there are some constants C_1, C_2 such that $\mathrm{E}_x|h(Y_0) - \bar{g}(\mu)| \leq C_1 x^{r/(2+\delta)} + C_2$ for all x large enough. Since $h(0) = 0$ and h is monotonic increasing, for $x > M$, by [29, eq. 23,],

$$\begin{aligned}\mathrm{E}_x|h(Y_0) - \bar{g}(\mu)| &= \mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu) + h(Y_0\mathbf{1}_{Y_0<0})| \\ &\leq \mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu)| + C.\end{aligned}$$

Using the Markov inequality stated in [29, eq. 14], for any fixed $\varepsilon \in (0,1)$ and $x > M$,

$$\begin{aligned}&\mathrm{E}_x[|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu)|\mathbf{1}_{Y_0\leq(1-\varepsilon)\mu}] && \text{(A-4)}\\ &\leq \mathrm{E}_x|\bar{g}(\mu)\mathbf{1}_{Y_0\leq(1-\varepsilon)\mu}| + \mathrm{E}_x|h(Y_0\mathbf{1}_{0<Y_0\leq(1-\varepsilon)\mu})| \\ &\leq \bar{g}(\mu)\mathrm{P}_x(Y_0 \leq (1-\varepsilon)\mu) + \mathrm{E}_x[h(\mu)\mathbf{1}_{Y_0\leq(1-\varepsilon)\mu})| \\ &\leq \frac{\bar{g}(\mu)(C_1\mu^r + C_2)}{\varepsilon^{2+\delta}\mu^{2+\delta}} + \frac{h(\mu)(C_1\mu^r + C_2)}{\varepsilon^{2+\delta}\mu^{2+\delta}}\,. && \text{(A-5)}\end{aligned}$$

If $\bar{g} \equiv h \neq g$, equation (A-5) reduces to $Ch(\mu)/\mu^{2+\delta-r}$. Recall that for $y > 0$, $a_0 + a_1 y \geq h(y)$, so that $a_0 + a_1\mu \geq h(\mu)$. Hence, $\mu \geq (h(\mu) - a_0)/a_1$ and (A-4) is bounded by: $Ch(\mu)/[h(\mu) - a_0]^{2+\delta-r} = C\tilde{h}(x)/[\tilde{h}(x) - a_0]^{2+\delta-r}$ which converges to 0 as $x \to \infty$. ($\tilde{h}(\cdot) = h(g^{-1}(\cdot)) = (h \circ g^{-1})(\cdot)$ is an increasing function, since it is a composition of increasing functions, and is therefore bounded by a constant, for $x > M$. If $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]$, it can be showed that $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0)]$. As $\sigma(X_0) \subseteq \mathcal{F}_{-1}$, for the tower property $\mathrm{E}_x[h(Y_0)] = \mathrm{E}[h(Y_0)|X_0] = \mathrm{E}[\mathrm{E}[h(Y_0)|\mathcal{F}_{-1}]|X_0] = \mathrm{E}[\bar{g}(\mu)|x] = \bar{g}(\mu)$. Moreover, we notice that $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0)] \leq h[\mathrm{E}_x(Y_0)] = h(\mu)$. Consequently, the above bound applies here. If $\bar{g} \equiv g \neq h$ we define (A-5) as $g(\mu)(C_1\mu^r + C_2)/\varepsilon^{2+\delta}\mu^{2+\delta} + h(\mu)(C_1\mu^r + C_2)/\varepsilon^{2+\delta}\mu^{2+\delta} = Cx/\mu^{2+\delta-r} + Ch(\mu)/\mu^{2+\delta-r}$ and it is bounded by $Cx/[x-a_0]^{2+\delta-r} + Ch(\mu)/[h(\mu)-a_0]^{2+\delta-r} = Cx/[x-a_0]^{2+\delta-r} + C\tilde{h}(x)/[\tilde{h}(x)-a_0]^{2+\delta-r}$, which converges to 0 as $x \to \infty$. It only remains to show that

$$\mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu)|\mathbf{1}_{Y_0>(1-\varepsilon)\mu} = \mathrm{E}_x|h(Y_0) - \bar{g}(\mu)|\mathbf{1}_{Y_0>(1-\varepsilon)\mu} \tag{A-6}$$

is "small". When $\bar{g} \equiv h$, this is straightforward by replacing $g(\cdot)$ by $h(\cdot)$ in [29, p. 826], establishing the existence of constants C_1, C_2, such that $\mathrm{E}_x|h(Y_0)-\bar{g}(\mu)| \leq C_1 x^{r/(2+\delta)} + C_2$, for all x large enough. For $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0)]$, the expectation (A-6) is bounded by $\mathrm{E}_x|h(Y_0)|\mathbf{1}_{Y_0>(1-\varepsilon)\mu} + \mathrm{E}_x|\bar{g}(\mu)|\mathbf{1}_{Y_0>(1-\varepsilon)\mu} \leq 2\bar{g}(\mu) \leq 2h(\mu)$ which is itself bounded by $2a_0 + 2a_1\mu \leq C_2 + C_1\mathrm{E}_x|Y_0| \leq C_2 + C_1\mu^{r/(2+\delta)} \leq C_2 + C_1 x^{r/(2+\delta)}$, for the concavity of $h(\cdot)$, for $\mu > 0$ when $x > M$, [29, p. 824], since $\mu \leq x/[b_1(x)(1-\varepsilon)]$ by equation (A-1) where $b_1(x)$ is bounded for $x > M$. Then, $\mathrm{E}_x|h(Y_0)-\bar{g}(\mu)| \leq C_1 x^{r/(2+\delta)} + C_2$, also for $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0)]$. Combining this result with (A-2) and (A-3), we have that, for all x enough large,

$$\mathrm{E}_x V(X_1) \leq C_2 + |\phi|x + |\theta|C_1 x^{r/(2+\delta)} + |\gamma|x \leq C + (|\phi| + |\gamma| + \varepsilon)x;$$

this gives the final result. For any $\varepsilon \in (0,1)$ there is some constant $G_3 < \infty$ such that for M large enough, $\mathrm{E}_x V(X_1) \leq (|\phi| + |\gamma| + \varepsilon)V(x) + G_3$.

When $x < -M$ and $\mu_t \in \mathbb{R}$, the previous proof holds directly by symmetry. In the other cases, according to the states discussed above, similar conditions are found; for convenience the results organized according to the domain of x and μ_t are reported below.

- For all $x \in A$, (*Cases 1-3*) There is some constant $G(M) < \infty$ such that $\mathrm{E}_x V(X_1) \leq G(M)$.
- For all $x > M$,
  - (*Cases 1-2*) For any $\varepsilon \in (0,1)$ there is some constant $G_3 < \infty$ such that for M large enough, $\mathrm{E}_x V(X_1) \leq (|\phi| + |\gamma| + \varepsilon)V(x) + G_3$.
  - (*Case 3*)
    - If $\bar{g}(\mu) \neq g(\mu)$ and $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0^*)]$ or $\bar{g} \equiv h$, there is some constant $U_3 < \infty$ such that $\mathrm{E}_x V(X_1) \leq |\gamma|V(x) + U_3$ for all $x > M$.
    - If $\bar{g}(\mu) = g(\mu)$, there is some constant $W_3 < \infty$ such that $\mathrm{E}_x V(X_1) \leq |\gamma - \theta|V(x) + W_3$ for all $x > M$.
- For all $x < -M$,
  - (*Case 1*) For any $\varepsilon \in (0,1)$ there is some constant $G_2 < \infty$ such that for M large enough, $\mathrm{E}_x V(X_1) \leq (|\phi| + |\gamma| + \varepsilon)V(x) + G_2$.
  - (*Cases 2-3*)
    - If $\bar{g}(\mu) \neq g(\mu)$ and $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0^*)]$ or $\bar{g} \equiv h$, there is some constant $U_2 < \infty$ such that $\mathrm{E}_x V(X_1) \leq |\gamma|V(x) + U_2$ for all $x < -M$.
    - If $\bar{g}(\mu) = g(\mu)$, there is some constant $W_2 < \infty$ such that $\mathrm{E}_x V(X_1) \leq |\gamma - \theta|V(x) + W_2$ for all $x < -M$.

These conditions can be combined to find the overall drift condition for all $x \in \mathbb{R}$, as follows. Consider *Case 2*; the other two cases are analogous. If $x < -M$ and $\bar{g}(\mu) = g(\mu)$, since $\varepsilon > 0$, we can write $\mathrm{E}_x V(X_1) \leq |\gamma - \theta|V(x) + W_2 \leq (|\gamma - \theta| + \varepsilon)V(x) + W_2$; where $x > M$, for M large enough, $\mathrm{E}_x V(X_1) \leq (|\phi| + |\gamma| + \varepsilon)V(x) + G_3$. Set $\xi = (|\phi| + |\gamma|) \vee |\gamma - \theta|$, then we can write $\mathrm{E}_x V(X_1) \leq$

$(\xi+\varepsilon)V(x)+\max\{W_2,G_3\}$. For $\varepsilon=(1-\xi)/2$, define $\eta=\xi+\varepsilon=\frac{\xi+1}{2}$, and choose M large enough. Then, for any $x\notin A$, we have $\mathrm{E}_xV(X_1)\leq\eta V(x)+L$, establishing the drift condition (A-1) for $|\gamma-\theta|+(|\phi|+|\gamma|)<1$. We remark that, although the range of V is $[0,\infty)$, we can easily replace V with $\tilde{V}(x)=|x|+1$ to get the range $[1,\infty)$. The same holds if $\bar{g}(\mu)=\mathrm{E}_x[h(Y_0^*)]$ or $\bar{g}\equiv h\neq g$, by setting $\eta=|\phi|+|\gamma|+\varepsilon$, giving the drift condition (A-1) for $|\phi|+|\gamma|<1$. □

Lemma 3 Assume (A3) holds. Then, the chain $\{X_t\}_{t\in T}$ defined in equation (14) is weak Feller.

Proof From (14) we have that $X_1(x)=\alpha+\phi h(Y_0^*(g^{-1}(x)))+\theta[h(Y_0^*(g^{-1}(x)))-\tilde{g}(x)]+\gamma x$. Since, by (A3), g^{-1} is continuous, $Y_0(g^{-1}(x))\Rightarrow Y_0(g^{-1}(x'))$ as $x\to x'$. The function Y_0^* that maps Y_0 to the domain of h is also continuous; then, $Y_0^*(g^{-1}(x))\Rightarrow Y_0^*(g^{-1}(x'))$ as $x\to x'$. For the same reason, we have that $h(Y_0^*(g^{-1}(x)))\Rightarrow h(Y_0^*(g^{-1}(x')))$ and $\tilde{g}(x)\Rightarrow\tilde{g}(x')$. So $X_1(x)\Rightarrow X_1(x')$ as $x\to x'$; this shows the weak Feller property. □

Lemma 4 Assume that Lemma 3, (A3) and (A4) hold. Then, $\{X_t\}_{t\in T}$ is asymptotic strong Feller.

Proof When $g\equiv\bar{g}$, it follows from equation (14) that $X_1(z)=\alpha+\phi h(Y_0^*(z))+\theta[h(Y_0^*(z))-\tilde{g}(z)]+\gamma z$. By Lemma 3, the chain is weak Feller, so if $h(Y_0^*(w))=h(Y_0^*(z))$, then $|X_1(z)-X_1(w)|=|-\theta(\tilde{g}(z)-\tilde{g}(w))+\gamma(z-w)|=|\gamma-\theta||z-w|$. From coupling theory, using [33, Prop. 3(g)] we can construct the random variables $g(Y_0^*(z))$ and $g(Y_0^*(w))$ in such a way that they have the marginal distributions π_z and π_w, and that $\mathrm{P}(g(Y_0^*(w))=g(Y_0^*(z)))=1-\|\pi_w(\cdot)-\pi_z(\cdot)\|_{TV}>1-B|z-w|$, where the inequality holds by assumption (A4). Note that $g(\cdot)$ and $h(\cdot)$ are one-to-one functions. Hence, we have $g(Y_0^*(w))=g(Y_0^*(z))\iff Y_0^*(w)=Y_0^*(z)\iff h(Y_0^*(w))=h(Y_0^*(z))$ (where $\iff$ means "if and only if"); so the conditional probability of $g(Y_0^*(w))=g(Y_0^*(z))$ or $h(Y_0^*(w))=h(Y_0^*(z))$ is equivalent. Therefore, the probability that the chains couple at $t=1$:

$$\mathrm{P}[g(Y_1^*(w))=g(Y_1^*(z))|h(Y_0^*(w))=h(Y_0^*(z))]>1-\left\|\pi_{X_1(z)}(\cdot)-\pi_{X_1(w)}(\cdot)\right\|_{TV} \tag{A-7}$$

which is bounded below by $1-B|\gamma-\theta||z-w|$. Then, the lower bound of the probability that the chains couple for all times $t=0,1,\ldots$ is obtained by iterating (A-7): $1-B|z-w|\sum_{t=0}^{\infty}(|\gamma-\theta|)^t=1-\frac{|z-w|B}{1-|\gamma-\theta|}$ where the equality holds by assumption (A3). The rest of the proof for the asymptotic strong Feller property follows as in [29, p. 819]. It is sufficient to replace $|\theta|$ by $|\gamma-\theta|$ anywhere. We omit the details. If $g\neq\bar{g}$ and $h(Y_0^*(w))=h(Y_0^*(z))$ we have $|X_1(z)-X_1(w)|=|-\theta(\tilde{g}(z)-\tilde{g}(w))+\gamma(z-w)|\leq|\theta||\tilde{g}(z)-\tilde{g}(w)|+|\gamma||z-w|$. Since $\tilde{g}(x)$ is Lipschitz with $L\leq1$, we obtain $|X_1(w)-X_1(z)|\leq(|\theta|+|\gamma|)|z-w|$. Hence, it can be immediately be seen that the proof for the former case ($\bar{g}\equiv g$) is also valid here by replacing $|\gamma-\theta|$ by $|\theta|+|\gamma|$. This completes the proof. □

Lemma 5 If (A3) holds, then a reachable point x_0 exists for the chain (14).

Proof Consider $\{X_t\}_{t\in T}$ where $X_t=g(\mu_t)$ and x_t is its sample counterpart. Firstly, consider the case in which $\bar{g}\equiv g$ and put, $h(0)=0$ (which simply

changes the value of the constant α). Equation (14) could be written as

$$x_t = \alpha + \gamma x_{t-1} + (\theta + \phi)h(Y^*_{t-1}) - \theta\tilde{g}(x_{t-1}). \tag{A-8}$$

Let us consider the case $Y^*_t = 0$, for $t = 1, \ldots, n$. Hence, by (A-8), $x_t = \alpha + (\gamma - \theta)x_{t-1}$. Then, set $x = \alpha/(1-\delta)$, where $\delta = \gamma - \theta$. Let $x \in \mathbb{R}$ and let C be an open set containing x. Then, by setting $x_0 = x$ and for all $t \geq 1$, $x_t = \alpha + \delta x_{t-1} = \alpha \sum_{j=0}^{t-1} \delta^j + \delta^t x_0$. Since $\delta \leq |\gamma - \theta| < 1$ for (A3), we have $\lim_{t\to\infty} x_t = x$ so that $\exists\, n \in \mathbb{N}$ such that $\forall t \geq n$, $x_t \in C$. For such n we have

$$\begin{aligned} P^n(x, C) &= \mathrm{P}_x(X_n \in C) \geq \mathrm{P}_x(X_n \in C, Y^*_0 = \cdots = Y^*_{n-1} = 0) \\ &= \mathrm{P}_x(X_n \in C | Y^*_0 = \cdots = Y^*_{n-1} = 0)\mathrm{P}_x(Y^*_0 = \cdots = Y^*_{n-1} = 0) \\ &= \mathrm{P}_x(Y^*_0 = \cdots = Y^*_{n-1} = 0) > 0. \end{aligned}$$

For the case $\bar{g}(\mu_t) = \mathrm{E}[h(Y^*_t)|\mathcal{F}_{t-1}]$, it can be immediately be seen that $\bar{g}(\mu_t) = 0$, for $t = 1, \ldots, n$ and (A-8) still holds, with γ instead of δ, as it follows by (A3) that $|\gamma| < 1$. When $\bar{g} \equiv h \neq g$ we consider the case $Y_t = c$, for $t = 1, \ldots, n$ so that $\mu_t = c$, for $t = 1, \ldots, n$ and $Y^*_t = c$, for $t = 1, \ldots, n$; and finally, set $h(c) = 0$ and (A-8) will be valid again, with γ instead of δ. □

A.2. Proof of Theorem 1

Theorem 1 follows directly from Lemmata 1-5. More precisely, if (A1)-(A2) and (A3) hold, the process $\{X_t\}_{t\in T}$ has at least a stationary distribution. The result is obtained by Lemmata 1-3 and Theorem 2 in [40]. Besides, if (A1)-(A4) hold, the stationary distribution of the process $\{X_t\}_{t\in T}$ is unique. This is immediate by Lemma 4, Lemma 5 and Theorem 3 in [29]. Finally, by Proposition 8 in [14], the stationarity of $\{Y_t\}_{t\in T}$ follows directly from the uniqueness of the stationary distribution of $\{X_t\}_{t\in T}$; this completes the proof. □

A.3. Proof of Corollary 1

Let us define $\nu_0 = \nu(\mu_0) = \nu(\mu) = \nu$ and set $g(\mu) = x$. It is worth noting that $\mathrm{E}_x\left[\frac{h(Y_0)-\tilde{g}(x)}{\nu}\right] = \frac{\mathrm{E}_x[h(Y^*_0)-\tilde{g}(x)]}{\nu}$. In fact ν is the standard deviation $\sigma(\mu)$ of $h(Y_0)$, which is constant w.r.t x (and then w.r.t μ). For this reason when $x \in A$ and $\mu \in \mathbb{R}$ (*Case 1*), the result of Lemma 2 holds here unchanged. When we have $x > M$; if ν is increasing w.r.t μ we have that as $x \to \infty$ ($\mu \to \infty$) ν goes to infinity as well (and $1/\nu \to 0$, then it is therefore bounded for $x > M$) or converges to a specific constant. In both cases the proof of Lemma 2 still holds with a modification of the constants C. The same thing (with signs inverted) holds as $x < -M$, provided that ν is decreasing w.r.t μ. *Case 2*, when $x \in [-M, \infty)$, holds as above, by setting without loss of generality, that $h(c) = 0$, since replacing $h(y)$ with $h(y) - h(c)$ simply changes the value of α; the same assumptions hold for g. When $x < -M$, we have $0 < \mu = g^{-1}(x) <$

$g^{-1}(0) = c$, ν is only required to be monotone w.r.t μ, indeed if it is decreasing $\sigma(\mu) > \sigma(c) = \xi$, instead, if it is increasing $\sigma(\mu) > \sigma(0) = \xi$, and then

$$\begin{aligned} \mathrm{E}_x V(X_1) &\leq C + (|\phi| + |\theta|/\nu) a_1 \mathrm{E}_x[Y_0 \mathbf{1}_{Y_0 \geq c}] + |\theta|/\nu |\tilde{g}(x)| + |\gamma||x| \\ &\leq C_2 + C_1/\nu\mu + |\theta|/\nu |\tilde{g}(x)| + |\gamma||x| \\ &\leq C_2 + C_1/\xi c + |\theta|/\xi |\tilde{g}(x)| + |\gamma||x| \\ &\leq C^* + |\theta|/\xi (C_1^* + C_2^* c^{r/(2+\delta)}) + |\gamma||x| \end{aligned}$$

which provide the same stationarity condition obtained in absence of the scaling sequence. For *Case 3* we have $0 < \mu < a$, also ν is required to be monotone, if it is increasing $\sigma(\mu) > \sigma(0) = \delta$, by contrast, if it is decreasing $\sigma(\mu) > \sigma(a) = \delta$, then

$$\mathrm{E}_x V(X_1) \leq C + (|\phi| + |\theta|/\delta) h(a - c) + |\theta|/\delta |\tilde{g}(x)| + |\gamma||x| \leq C + |\theta|/\nu h(a - c) + |\gamma||x|$$

which again provide the same stationarity condition. Then, Lemma 2 holds also for the chain (15).

As far as the Feller properties are concerned, it is easy to see that the weak Feller condition is satisfied since, in general, $\sigma^2(\mu)$ is continuous for μ (and then for x). Hence, Lemma 3 holds. Also, in order to prove Theorem 1, the asymptotic strong Feller property remains to be verified. Define $\tilde{Y}_0 = h(Y_0)$ and $\tilde{\mu} = \bar{g}(\mu)$. We compute the scaling sequence from the first order Taylor expansion: $b(\tilde{Y}_0) \approx b(\tilde{\mu}) + b'(\tilde{\mu})(\tilde{Y}_0 - \tilde{\mu})$ so as to obtain $\mathrm{V}[b(\tilde{Y}_0)] \approx b'(\tilde{\mu})^2 \nu^2$ where here $\nu^2 = \mathrm{V}[h(Y_0)]$. The function b is selected as Lipschitz with constant not greater than 1. Then, by using the variance stabilizing transformation (VST) we obtain a constant variance c^2 w.r.t. the mean $\tilde{\mu}$. After that, we take the approximation $\frac{h(Y_0) - \bar{g}(\mu)}{\nu} \approx \frac{b(\tilde{Y}_0) - b(\tilde{\mu})}{c}$ and show the asymptotic strong Feller property on this approximated version. The remaining part of the proof is the same as Lemma 4. We omit the details. Finally, as we are in the case where $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]$ here, the existence of a reachable point does not require any modification of the proof for Lemma 5. Hence, for (15), Corollary 1 holds. □

A.4. Proof of Theorem 2

Equation (20) may be rewritten in the following way. For the mean-value theorem, $\tilde{g}(x_s) - \tilde{g}(0) = \tilde{g}'(u_s) x_s = c_s x_s$ for $s = 0, \ldots, t$ and $0 < u_s < x_s$. We can replace $\tilde{g}(x)$ with $\tilde{g}(x) - \tilde{g}(0)$; this simply changes the value of the constant α with $\alpha - \theta\tilde{g}(0)$. Then, set

$$g^{\rho}_{y_1}(x) = \alpha + \gamma x + (\phi + \theta) h(y_0) - \theta \tilde{g}(x) = \alpha + \delta h(y_0) + r_0 x \tag{A-9}$$

where $\delta = \phi + \theta$, $r_0 = \gamma - \theta c_0$ and $x_0 = x$. Then, for $s \leq t$, by using IRF, we have,

$$g^{\rho}\langle y_{s:t} \rangle(x) = \alpha \sum_{j=0}^{t-s} \prod_{i=0}^{j-1} r_{t-i} + \delta \sum_{j=0}^{t-s} \prod_{i=0}^{j-1} r_{t-i} h(y^*_{t-j}) + \prod_{j=0}^{t-s} r_j x \,, \tag{A-10}$$

where $r_{t-i} = 1$ for $i = -1$. Moreover, from (A-10), and by equation (19), we can define $g^\rho\langle Y_{-\infty:t}\rangle \coloneqq \alpha \sum_{j=0}^{\infty} \prod_{i=0}^{j-1} r_{t-i} + \delta \sum_{j=0}^{\infty} \prod_{i=0}^{j-1} r_{t-i} h(Y^*_{t-j})$. The proof is carried out specifically for $\bar{g}(\cdot) \neq g(\cdot)$. It is worth noting that $\left|\sup_j \{c_j\}\right| \leq 1$ for the Lipschitz continuity of $\tilde{g}$. Then, from Theorem 1, we have $0 < r_- \leq |r_j| \leq |\gamma| + |\theta c_j| \leq |\gamma| + |\theta| \leq \tilde{r} < 1$ where $r_- = \min(r_j)$. However, one can immediately see that (A-9) also holds in the simpler case $\bar{g}(\cdot) = g(\cdot)$, with $r_0 = r = \gamma - \theta$, where $|\gamma - \theta| < 1$ from Theorem 1. Let $\{Y_n\}_{n\in\mathbb{Z}}$ be a strictly stationary and ergodic process, satisfying Theorem 1. The proof of Theorem 2 holds if assumptions (B1)-(B3) in [14, Thr. 19] are verified. Assumptions (B1) and (B2) hold in our case for the stationarity of Y_t and the continuity of $g^\rho_y(x)$ w.r.t. ρ and $q(\cdot; y)$ w.r.t. x. Hence, the estimator $\hat{\rho}_{n,x}$ is well-defined. Assumption (B3)-(iii) holds here for the discreteness of Y_t, see [14, Rmk. 18]. This condition is required in order to obtain a solvable maximization problem. It remains to show (B3)-(i) and (B3)-(ii). (B3)-(i): $\lim_{m\to\infty} \sup_{\rho\in\Lambda} |g^\rho\langle Y_{-m:0}\rangle(x) - g^\rho\langle Y_{-\infty:0}\rangle| = 0$, a.s., which ensures that, regardless of the initial value of $X_{-m} = x$, X_0 (and thus X_t) can be approximated by a quantity involving the infinite past of the observations. (B3)-(ii): $\lim_{t\to\infty} \sup_{\rho\in\Lambda} |\log q(g^\rho\langle Y_{1:t-1}\rangle(x); Y_t) - \log q(g^\rho\langle Y_{-\infty:t-1}\rangle; Y_t)| = 0$, a.s., with the first element $\log q(g^\rho\langle Y_{1:t-1}\rangle(x); Y_t) = Y_t g^\rho\langle Y_{1:t-1}\rangle(x) - A[g^\rho\langle Y_{1:t-1}\rangle(x)] + d(Y_t)$, the second element is defined as $\log q(g^\rho\langle Y_{-\infty:t-1}\rangle; Y_t) = Y_t g^\rho\langle Y_{-\infty:t-1}\rangle - A[g^\rho\langle Y_{-\infty:t-1}\rangle] + d(Y_t)$. Intuitively, this assumption allows the conditional log-likelihood function to be approximated by a stationary sequence. In order to prove (B3)-(i) note that, a.s.

$$\sup_{\rho\in\Lambda} |g^\rho\langle Y_{-\infty:0}\rangle| \leq \frac{\tilde{\alpha}}{1-\tilde{r}} + \tilde{\delta} \sum_{j=0}^{\infty} \tilde{r}^j |h(Y^*_{-j})| = \hat{g}\langle Y_{-\infty:0}\rangle\,, \tag{A-11}$$

which has finite expectation, and then is finite according to (H1). In fact, $h(Y^*_t)$ is stationary and $|h(Y_0)| \leq a_0 + a_1|Y_0|$, for Case 1. For Case 2, $h(Y^*_0) \leq a_1 Y^*_0$ and $\mathrm{E}[Y^*_0] \leq \mathrm{E}[Y_0] + c$. In Case 3 $h(\cdot)$ and Y_t are bounded so their expectations are finite. It holds also that

$$|g^\rho\langle Y_{-\infty:t-1}\rangle| \leq \frac{\tilde{\alpha}}{1-\tilde{r}} + \tilde{\delta} \sum_{j=0}^{\infty} \tilde{r}^j |h(Y^*_{t-1-j})| \tag{A-12}$$

$$|g^\rho\langle Y_{1:t-1}\rangle(x)| \leq \tilde{\alpha} \sum_{j=0}^{t-2} \tilde{r}^j + \tilde{\delta} \sum_{j=0}^{t-2} \tilde{r}^j |h(Y^*_{t-1-j})| + \tilde{r}^{t-1}|x| \tag{A-13}$$

which has a finite expectation by (H1). Let $d_1 = |g^\rho\langle Y_{-m:0}\rangle(x) - g^\rho\langle Y_{-\infty:0}\rangle|$ and $j = m + l + 1$. Then,

$$\begin{aligned} d_1 &= \left| \alpha \sum_{l=0}^{\infty} \prod_{i=0}^{m+l} r_{-i} + \delta \sum_{l=0}^{\infty} \prod_{i=0}^{m+l} r_{-i} h(Y^*_{-m-l-1}) + \prod_{j=0}^{m} r_j x \right| \\ &\leq \left| \prod_{i=0}^{m} r_{-i} \right| \left| \alpha \sum_{l=0}^{\infty} \prod_{i=m+1}^{m+l+1} r_{-i} + \delta \sum_{l=0}^{\infty} \prod_{i=m+1}^{m+l+1} r_{-i} h(Y^*_{-m-l-1}) \right| + \left| \prod_{j=0}^{m} r_j x \right| \end{aligned}$$

$$\leq \tilde{r}^{m+1}\left(\tilde{\alpha}\sum_{l=0}^{\infty}\tilde{r}^{l}+\tilde{\delta}\sum_{l=0}^{\infty}\tilde{r}^{l}|h(Y^{*}_{-m-l-1})|+|x|\right)$$

converges to 0 as $m\to\infty$ by (H1) and [14, Lem. 34]. Thus (B3)-(i) holds. We now move to (B3)-(ii), $\sup_{\rho\in\Lambda}|\log q(g^{\rho}\langle Y_{1:t-1}\rangle(x);Y_t)-\log q(g^{\rho}\langle Y_{-\infty:t-1}\rangle;Y_t)|$, which is bounded by the sum of $Y_t\sup_{\rho\in\Lambda}|f[g^{\rho}\langle Y_{1:t-1}\rangle(x)]-f[g^{\rho}\langle Y_{-\infty:t-1}\rangle]|$ and $\sup_{\rho\in\Lambda}|A[g^{\rho}\langle Y_{1:t-1}\rangle(x)]-A[g^{\rho}\langle Y_{-\infty:t-1}\rangle]|$. Consider

$$\begin{aligned}|g^{\rho}\langle Y_{1:t-1}\rangle(x)-g^{\rho}\langle Y_{-\infty:t-1}\rangle|&\leq\tilde{r}^{t-1}\left(\tilde{\alpha}\sum_{l=0}^{\infty}\tilde{r}^{l}+\tilde{\delta}\sum_{l=0}^{\infty}\tilde{r}^{l-}|h(Y^{*}_{-l})|+|x|\right)\\&=\tilde{r}^{t-1}\left(|x|+\hat{g}\langle Y_{-\infty:0}\rangle\right)\end{aligned}$$

for (A-11), and for $l=j$ when $t-1=0$. This implies that

$$Y_t\sup_{\rho\in\Lambda}|g^{\rho}\langle Y_{1:t-1}\rangle(x)-g^{\rho}\langle Y_{-\infty:t-1}\rangle|\leq Y_t\tilde{r}^{t-1}\left(|x|+\hat{g}\langle Y_{-\infty:0}\rangle\right)\xrightarrow{t\to\infty}0\ \text{a.s.}$$

according to (A-11) and by [14, Lem. 34], under (H1). Now, for the mean value theorem, $\sup_{\rho\in\Lambda}|A[g^{\rho}\langle Y_{1:t-1}\rangle(x)]-A[g^{\rho}\langle Y_{-\infty:t-1}\rangle]|$ is bounded by

$$\begin{aligned}&\sup_{\rho\in\Lambda}|A'(C_{t-1})|\,|g^{\rho}\langle Y_{1:t-1}\rangle(x)-g^{\rho}\langle Y_{-\infty:t-1}\rangle|\\&\leq\sup_{\rho\in\Lambda}|A'(C_{t-1})|\,\tilde{r}^{t-1}\left(|x|+\hat{g}\langle Y_{-\infty:0}\rangle\right)\end{aligned}\tag{A-14}$$

as $\min\{g^{\rho}\langle Y_{1:t-1}\rangle(x),g^{\rho}\langle Y_{-\infty:t-1}\rangle\}\leq C_{t-1}\leq\max\{g^{\rho}\langle Y_{1:t-1}\rangle(x),g^{\rho}\langle Y_{-\infty:t-1}\rangle\}$ and he function (A-14) tends to 0 as $t\to\infty$, for [14, Lem. 34] and the finiteness of $\mathrm{E}[(\log|A'(C_{t-1})|)_{+}]$, which is true for (H1). The same argument of (A-14) hold with $f(\cdot)$ instead of $A(\cdot)$, and the details are omitted. Then, (B3)-(ii) holds, and this completes the proof. □

A.5. Proof of Theorem 3

Note that $P(x,A)=\int_A q(x;y)\mu(dy)$. By the stationarity of Y_t and (H1), Theorem 2 holds. It remains to prove that $\mathrm{P}_{\star}=\{\rho_{\star}\}$, where $\rho_{\star}=(\alpha_{\star},\gamma_{\star},\phi_{\star},\theta_{\star})$. This follows from [14, Prop. 21], once we have shown that

(LP1) $X_0=g^{\rho_{\star}}\langle Y_{-\infty:0}\rangle$, a.s.
(LP2) $x\mapsto P(x;\cdot)$ is one-to-one, i.e, if $P(x;\cdot)=P(x';\cdot)$ implies that $x=x'$.
(LP3) $g^{\rho_{\star}}\langle Y_{-\infty:0}\rangle=g^{\rho}\langle Y_{-\infty:0}\rangle$ a.s. implies that $\rho=\rho_{\star}$.

Consider, for $m\geq 0$,

$$g^{\rho_{\star}}\langle Y_{-m:0}\rangle(X_{-m-1})=\alpha_{\star}\sum_{j=0}^{m}\prod_{i=0}^{j-1}r_{\star-i}+\delta_{\star}\sum_{j=0}^{m}\prod_{i=0}^{j-1}r_{\star-i}h(Y^{*}_{-j})+\prod_{j=0}^{m}r_{\star j}X_{-m-1}\,.$$

For $m\to\infty$ we have $\prod_{j=0}^{m}r_{\star j}X_{-m-1}\to 0$ in fact $\sup_j\{r_{\star j}\}=r^{*}\leq\tilde{r}<1$. Hence, $X_0=\lim_{m\to\infty}g^{\rho_{\star}}\langle Y_{-m:0}\rangle(X_{-m-1})=g^{\rho_{\star}}\langle Y_{-\infty:0}\rangle$, a.s. thus (LP1) holds.

Moreover, (LP2) holds as well because $P(x;\cdot)$ is the cumulative distribution function of $q(x;\cdot)$, which is the exponential family of parameter $\mu = g^{-1}(x)$. It remains to check (LP3). Consider $g^{\rho_\star}\langle Y_{-\infty:0}\rangle - g^{\rho}\langle Y_{-\infty:0}\rangle$ equals

$$\sum_{j=0}^{\infty}\prod_{i=0}^{j-1}(\alpha_\star\gamma_\star - \alpha\gamma) + \sum_{j=0}^{\infty}\prod_{i=0}^{j-1}(\phi_\star\gamma_\star + \theta_\star\gamma_\star - \phi\gamma - \theta\gamma)\, h(Y^*_{-j})$$
$$+ \sum_{j=0}^{\infty}\prod_{i=0}^{j-1}(\alpha\theta - \alpha_\star\theta_\star)\, c_{-i} + \sum_{j=0}^{\infty}\prod_{i=0}^{j-1}\left(\phi\theta + \theta^2 - \phi_\star\theta_\star - \theta_\star^2\right) c_{-i} h(Y^*_{-j})$$

where $\delta_\star = \phi_\star + \theta_\star$, $r_{\star s} = \gamma_\star - \theta_\star c_s$ for $-j+1 \le s \le 0$. Clearly, only if $\alpha = \alpha_\star, \gamma = \gamma_\star, \theta = \theta_\star, \phi = \phi_\star$ (so $\rho = \rho_\star$), we have $g^{\rho_\star}\langle Y_{-\infty:0}\rangle - g^{\rho}\langle Y_{-\infty:0}\rangle = 0$, which completes the proof. □

A.6. Proof of Theorem 4

The proof of the theorem is based on [15, Thm. 4.2], and requires to prove that all the assumptions therein, (A1), (A4), (A5) and (A7), hold when the assumptions of Theorem 4 hold. First of all, note that (A1) is satisfied for the stationarity of Y_t and (A4) is assumed in Theorem 4. Moreover, (A5) follows by $\mu = A'(x_\star)$. It remains to prove assumption (A7). Let $g^\bullet\langle Y_{-\infty:t-1}\rangle : \rho \mapsto g^\rho\langle Y_{-\infty:t-1}\rangle$ and $g^\bullet\langle Y_{1:t-1}\rangle(x) : \rho \mapsto g^\rho\langle Y_{1:t-1}(x)\rangle$. We assume that the function $x \mapsto q(x,y)$ is twice differentiable. For all twice differentiable $x_t : \mathrm{P} \to \mathbb{R}$ and all $y \in \mathbb{R}$, define the score function $\chi^\rho(x_t(\rho), y_t) = \nabla_\rho x_t(\rho)\frac{\partial \log q(x_t,y_t)}{\partial x_t}$ and the Hessian matrix $K^\rho(x_t(\rho), y_t) = \nabla^2_\rho x_t(\rho)\frac{\partial \log q(x_t,y_t)}{\partial x_t} + \nabla_\rho x_t(\rho)\nabla_\rho x_t(\rho)'\frac{\partial^2 \log q(x_t,y_t)}{\partial x_t^2}$. In order to prove asymptotic normality for the QMLE (21) by following the line of [15] the following assumptions are required to hold true. Define $\|X\|$ any suitable norm for the object X.

(A7): $\forall y \in \mathbb{R}$, the function $x \mapsto q(x,y)$ is twice continuously differentiable. Moreover, there exists $\epsilon > 0$ and a family of P-a.s. finite random variables $g^\rho\langle Y_{-\infty:t}\rangle$, for $(\rho,t) \in \mathrm{P}\times\mathbb{Z}$, such that $g^{\rho_\star}\langle Y_{-\infty:0}\rangle$ is in the interior of S, the function $\rho \mapsto g^\rho\langle Y_{-\infty:0}\rangle$ is, P-a.s., twice continuously differentiable on some ball $B(\rho_\star,\epsilon)$ and for all $x \in S$, almost surely

(i) $\lim_{t\to\infty} \left\|\chi^{\rho_\star}\left(g^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right) - \chi^{\rho_\star}\left(g^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t\right)\right\| = 0,$
(ii) $\lim_{t\to\infty} \sup_{\rho\in B(\rho_\star,\epsilon)} \left\|K^{\rho}\left(g^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right) - K^{\rho}\left(g^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t\right)\right\| = 0,$
(iii) $\mathrm{E}\left[\left\|\chi^{\rho_\star}\left(g^\bullet\langle Y_{-\infty:0}\rangle, Y_1\right)\right\|^2\right], \mathrm{E}\left[\sup_{\rho\in B(\rho_\star,\epsilon)}\left\|K^{\rho}\left(g^\bullet\langle Y_{-\infty:0}\rangle, Y_1\right)\right\|\right] < +\infty.$

Intuitively, (A7) implies that the score function and the information matrix of the data can be approximated by the infinite past of the process. We start from (A7)-(i). Clearly $\lim_{t\to\infty}\|\mathbf{a}-\mathbf{b}\| = 0$ holds if $\lim_{t\to\infty}|a_j - b_j| = 0$ for all j. Put $\chi^\rho(\cdot,\cdot) = \left[\chi^\alpha(\cdot,\cdot), \chi^\phi(\cdot,\cdot), \chi^\gamma(\cdot,\cdot), \chi^\theta(\cdot,\cdot)\right]'$. Consider the derivatives of the (quasi) log-likelihood, say $\chi^{\rho_\star}\left(g^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right)$, as

$$\left[Y_t f'\left[g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right] - A'\left[g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right]\right]\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial \rho_\star}$$

where, given that $r_j = \gamma - \theta c_j$,for the product rule, $\partial_1 = \partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)/\partial\gamma_\star$. Then,

$$\partial_1 = \sum_{j=0}^{t-2}\prod_{i=0}^{j-1} r_{t-1-i}\sum_{i=0}^{j-1}\frac{1}{r_{t-1-i}}\left[\alpha_\star + (\phi_\star+\theta_\star)h(Y^*_{t-1-j})\right] + \prod_{j=0}^{t-2} r_j x \sum_{i=0}^{t-2}\frac{1}{r_i}$$

where we have made implicit $r^\star_j = \gamma_\star - \theta_\star c_j = r_j$ to avoid excesses in the notation. With trivial manipulations it follows that

$$\begin{aligned}
&|\chi^{\gamma_\star}\left(g^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right) - \chi^{\gamma_\star}\left(g^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t\right)| \\
&= |Y_t|\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right| |f'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right] - f'\left[g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right]| \\
&\quad + |Y_t|\,|f'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right]|\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star} - \frac{\partial g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle}{\partial\gamma_\star}\right| \\
&\quad + \left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right| |A'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right] - A'\left[g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right]| \qquad \text{(A-15)} \\
&\quad + |A'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right]|\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star} - \frac{\partial g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle}{\partial\gamma_\star}\right|. \qquad \text{(A-16)}
\end{aligned}$$

It is possible to verify that

$$\begin{aligned}
\left|\frac{\partial g^{\rho_\star}\langle Y_{-\infty:0}\rangle}{\partial\gamma}\right| &\le |\alpha|\sum_{j=0}^{\infty}\tilde{r}^j\sum_{i=0}^{j-1}\frac{1}{r_-} + |\phi+\theta|\sum_{j=0}^{\infty}\tilde{r}^j\left|h(Y^*_{-j})\right|\sum_{i=0}^{j-1}\frac{1}{r_-} \\
&= \tilde{\alpha}\sum_{j=0}^{\infty}\frac{\tilde{r}^j}{r_-}j + \tilde{\delta}\sum_{j=0}^{\infty}\frac{\tilde{r}^j}{r_-}j\left|h(Y^*_{-j})\right| = \frac{\partial\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\gamma} < \infty \qquad \text{(A-17)}
\end{aligned}$$

which is finite for (H2). For the same argument

$$\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle}{\partial\gamma}\right| \le \frac{\partial\hat{g}\langle Y_{1:t-1}\rangle}{\partial\gamma} < \infty. \qquad \text{(A-18)}$$

Now the difference $\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star} - \frac{\partial g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle}{\partial\gamma_\star}\right|$ is bounded by

$$\begin{aligned}
&|\alpha_\star|\sum_{l=0}^{\infty}\frac{\tilde{r}^{t+l-1}}{r_-}(t+l-1) + \\
&+ |\phi_\star+\theta_\star|\sum_{l=0}^{\infty}\frac{\tilde{r}^{t+l-1}}{r_-}(t+l-1)\left|h(Y^*_l)\right| + \frac{\tilde{r}^{t-1}}{r_-}(t-1)\left|x\right| \\
&\le \tilde{r}^{t-1}\left(\tilde{\alpha}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-}l + \tilde{\delta}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-}l\left|h(Y^*_{-l})\right|\right) + \\
&\tilde{r}^{t-1}(t-1)\left(\tilde{\alpha}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-} + \tilde{\delta}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-}\left|h(Y^*_{-l})\right| + \frac{|x|}{r_-}\right)
\end{aligned}$$

$$= \tilde{r}^{t-1}\frac{\partial \hat{g}\langle Y_{-\infty:0}\rangle}{\partial \gamma} + \tilde{r}^{t-1}(t-1)\left(\frac{\hat{g}\langle Y_{-\infty:0}\rangle}{r_-} + \frac{|x|}{r_-}\right) \xrightarrow{t\to\infty} 0$$

almost surely, so that (A-16) tends to 0 as $t \to \infty$ according to [14, Lem. 34], (H1) and equation (A-17). The mean value theorem allows to rewrite equation (A-15) as $\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial \gamma_\star}\right| |A''(C_{t-1})|\, |g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|$, which tends to 0 as $t \to \infty$ for the same reason in (A-14) if the following expectation is finite

$$\mathrm{E}\left(\log\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial \gamma_\star}\right|\right)_+ + \mathrm{E}\left(\log|A''(C_{t-1})|\right)_+ . \tag{A-19}$$

The first term of (A-19), $\mathrm{E}\left(\log\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial \gamma_\star}\right|\right)_+ \leq \mathrm{E}\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial \gamma_\star}\right| < \infty$ is finite, since, for (H2), the expectation of (A-18) is finite. The proof in the second term of (A-19) follows from the mean-value theorem. Denote $M = \mathrm{E}\left(\log|A'(g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle)|\right)_+ + \mathrm{E}\left(\log|A'(g^{\rho_\star}\langle Y_{1:t-1}\rangle(x))|\right)_+ + 1$, which is finite for (H1). We can rewrite $\mathrm{E}\left(\log|A''(C_{t-1})|\right)_+$ as

$$\begin{aligned}
&\mathrm{E}\left(\log\frac{|A'(g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle) - A'(g^{\rho_\star}\langle Y_{1:t-1}\rangle(x))|}{|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|}\right)_+ \\
&\leq M + \mathrm{E}\left(-\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|\right)_+ \\
&\leq M - \mathrm{E}\left(\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|\right)_- \\
&= M - \frac{1}{2}\mathrm{E}\left(|\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)||\right) + \\
&\quad + \frac{1}{2}\mathrm{E}\left(\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|\right) \\
&\leq M + \frac{1}{2}\mathrm{E}\,|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle| + \frac{1}{2}\mathrm{E}\,|g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|
\end{aligned} \tag{A-20}$$

which is finite as the expectations of (A-12) and (A-13) are for (H1). The same results of (A-15) and (A-16) apply similarly for $f'(\cdot)$, thus are omitted. Hence, (A7)-(i) is proved. We now move to (A7)-(ii). Consider $K^\rho\left(g^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right)$ as

$$\begin{aligned}
&\left[Y_t f'\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right] - A'\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right]\right]\frac{\partial^2 g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\rho\,\partial\rho'} + \\
&+ \frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\rho}\frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\rho'}\left[Y_t f''\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right] - A''\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right]\right].
\end{aligned}$$

The proof is shown for a single derivative, the proof of the others is immediate. The term $\left|K^\theta\left(g^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right) - K^\theta\left(g^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t\right)\right|$ is bounded by

$$\begin{aligned}
&\left[|Y_t|\,|f'\left(g^\rho\langle Y_{-\infty:t-1}\rangle\right)| + |A'\left(g^\rho\langle Y_{-\infty:t-1}\rangle\right)|\right] \\
&\quad\times\left|\frac{\partial^2 g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2} - \frac{\partial^2 g^\rho\langle Y_{-\infty:t-1}\rangle}{\partial\theta^2}\right|
\end{aligned} \tag{A-21}$$

$$+\left|\frac{\partial^2 g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2}\right|\left|A'\left[g^\rho\langle Y_{-\infty:t-1}\rangle\right]-A'\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right]\right| \tag{A-22}$$

$$+\left|\frac{\partial^2 g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2}\right|\left|Y_t\right|\left|f'\left[g^\rho\langle Y_{-\infty:t-1}\rangle\right]-f'\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right]\right| \tag{A-23}$$

$$+\left(\frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right)^2\left|A''\left[g^\rho\langle Y_{-\infty:t-1}\rangle\right]-A''\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right]\right| \tag{A-24}$$

$$+\left(\frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right)^2\left|Y_t\right|\left|f''\left[g^\rho\langle Y_{-\infty:t-1}\rangle\right]-f''\left[g^\rho\langle Y_{1:t-1}\rangle(x)\right]\right| \tag{A-25}$$

$$+\left[\left|Y_t\right|\left|f''\left(g^\rho\langle Y_{-\infty:t-1}\rangle\right)\right|+\left|A''\left(g^\rho\langle Y_{-\infty:t-1}\rangle\right)\right|\right]$$
$$\times\left|\left(\frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right)^2-\left(\frac{\partial g^\rho\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right)^2\right|.$$

By the definition of second derivative it can be easily shown that

$$\left|\frac{\partial^2 g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2}-\frac{\partial^2 g^\rho\langle Y_{-\infty:t-1}\rangle}{\partial\theta^2}\right|\leq 2\tilde{r}^{t-1}(t-1)^2\left(7\frac{\partial^2\hat{g}^\rho\langle Y_{-\infty:0}\rangle}{\partial\theta^2}+\frac{|x|}{r_-^2}\right)$$

which is finite as $\frac{\partial^2\hat{g}^\rho\langle Y_{-\infty:0}\rangle}{\partial\theta^2}=\tilde{\alpha}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-^2}l^2+\left(\tilde{\alpha}+\tilde{\phi}+1\right)\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-^2}l^2\left|h(Y_{-l}^*)\right|$ has a finite expectation, according to (H1). So, the first element (A-21) tends to 0 as $t\to\infty$ for (H1), by [14, Lem. 34]. The same holds for the elements (A-22) and (A-23) since (A-19) is verified (the only difference here is that the expectation of the second derivative is required to be finite but $\mathrm{E}\left(\log\left|\frac{\partial^2 g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2}\right|\right)_+\leq \mathrm{E}\left|\frac{\partial^2 g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2}\right|<\infty$ always for (H1)). Equations (A-24) and (A-25) also tend to 0 as $t\to\infty$ because of [14, Lem. 34] and $\mathrm{E}\left(\log|A'''(C_{t-1})|\right)_+<\infty$, $\mathrm{E}\left(\log|f'''(C_{t-1})|\right)_+<\infty$; the proof is analogous to (A-20). Finally, it follows that the last element also tends to 0 as $t\to\infty$ for (H1), by [14, Lem. 34], because it can be rewritten as

$$\left|\frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right|\left|\frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta}-\frac{\partial g^\rho\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right|$$
$$+\left|\frac{\partial g^\rho\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right|\left|\frac{\partial g^\rho\langle Y_{1:t-1}\rangle(x)}{\partial\theta}-\frac{\partial g^\rho\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right|$$

completing the proof for (A7)-(ii). It remains to show (A7)-(iii): the score is bounded by

$$\left(Y_1^2 f'\left[g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right]^2+A'\left[g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right]^2\right)\sum_{i=1}^{4}\left(\frac{\partial\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\rho_i}\right)^2$$

which provides a finite expectation for the Hölder's inequality and condition (H2). An analogously result holds for the Hessian $\|K^\rho\left(g^\bullet\langle Y_{-\infty:0}\rangle, Y_1\right)\|$; detailed proofs can be found in [2, Ch. 3]; this completes the proof. □

A.7. Proof of equivalence of (A4) and (A5) for Negative Binomial

For $d_{TV}(g(Y_t^*(z)), g(Y_t^*(w))) = d_{TV}(Y_t(z), Y_t(w))$, the coupling inequality, as in [39], ensures that $d_{TV}(Y_t(z), Y_t(w)) \leq \mathrm{P}(Y_t(z) \neq Y_t(w))$. So, bounding $\mathrm{P}(Y_t(z) \neq Y_t(w))$ with a Lipschitz function is equivalent to proving Assumption (A4). Suppose that $z > w$ and let $Y_t(z) \sim NB(a, p_z = \frac{a}{g^{-1}(z)+a})$ and $Y_t(w) \sim NB(a, p_w = \frac{a}{g^{-1}(w)+a})$; set $Y_t(z) = U + Y_t(w)$, so $U = Y_t(z) - Y_t(w)$, and, by using the discrete-variable convolution, we have

$$\begin{aligned}\mathrm{P}(U = u) &= \sum_{k=0}^{\infty} \mathrm{P}(Y_t(w) = k)\mathrm{P}(Y_t(z) = k + u) \\ &= \sum_{k=0}^{\infty} \binom{a+k-1}{k} p_z^a (1-p_z)^k \binom{a+k+u-1}{k+u} p_w^a (1-p_w)^{k+u}\end{aligned}$$

and then

$$\mathrm{P}(U = 0) = (p_z p_w)^a \sum_{k=0}^{\infty} \binom{a+k-1}{k}^2 [(1-p_z)(1-p_w)]^k .$$

The coupling probability could be written as

$$\begin{aligned}\mathrm{P}(Y_t(z) \neq Y_t(w)) &= \mathrm{P}(U \neq 0) = 1 - \mathrm{P}(U = 0) \\ &\leq 1 - (p_z p_w)^a \sum_{k=0}^{\infty} \binom{a+k-1}{k} [(1-p_z)(1-p_w)]^k \\ &= 1 - \left(\frac{p_z p_w}{1-(1-p_z)(1-p_w)}\right)^a \\ &= 1 - \left(\frac{1}{1 + \frac{1-p_z}{p_z} + \frac{1-p_w}{p_w}}\right)^a = 1 - \left(\frac{1}{D}\right)^a \\ &= 1 - \left(\frac{g^{-1}(w) - g^{-1}(z)}{D\,(g^{-1}(w) - g^{-1}(z))}\right)^a \\ &\leq 1 - \left(-\frac{\zeta(z-w)}{D\,(g^{-1}(w) - g^{-1}(z))}\right)^a \qquad \text{(A-26)} \\ &= 1 - \left(\frac{\zeta(z-w)}{D\,(g^{-1}(z) - g^{-1}(w))}\right)^a \\ &\leq 1 - \left(\frac{\zeta(z-w)}{aD^*}\right)^a \qquad \text{(A-27)}\end{aligned}$$

where $D \geq 1$ and and $D\left(g^{-1}(z) - g^{-1}(w)\right) = D_1$. In equation (A-27) we put $D^* = \max\{D, D_1\}$. The inequality (A-26) holds because the function $g^{-1}(\cdot)$ is Lipschitz with constant ζ. Then, (A-27) is Lipschitz as well with constant ζ for $z \in [w, w + aD^*/\zeta]$, since the absolute value of its derivative is bounded by ζ, and this gives the desired result. □

A.8. Insights about conditions (H1)-(H2)

In this section, conditions (H1)-(H2) introduced in Section 4.1, are verified for particular cases of interest, with the aim of showing that (i) they hold for a large variety of models and (ii) they are easily verifiable. Of course, existence of moments of Y_t cannot be directly proved, as they rely on the unknown unconditional distribution of of Y_t. However, moments conditions are quite usual assumptions in the context of ML inference. We focus on other expectations. For convenience in terms of notation, in this paragraph we write $g^{\rho}\langle Y_{-\infty:t}\rangle = X_t$.

We start from the standard case in which the link $g(\cdot)$ is canonical; here the conditions on the derivative of $f(\cdot)$ hold automatically, since $f(X_t) = X_t$, $f'(X_t) = 1$ and $f''(X_t) = 0$, hence the respective expectations are finite. The moment condition for the derivatives of $A(\cdot)$ can be easily proved by noting that, from the properties of the exponential family, $A'(X_t) \equiv g^{-1}(X_t)$; in this case, the inverse of the link function is Lipschitz continuous in our case of interest; see Assumption (A5). Then, we can write $g^{-1}(X_t) - g^{-1}(0) \leq L|X_t|$ and

$$\begin{aligned}(\log|g^{-1}(X_t)|)_+ &= (\log|g^{-1}(X_t) - g^{-1}(0) + g^{-1}(0)|)_+ \\ &\leq \log^{\star}|g^{-1}(X_t) - g^{-1}(0)| + b \\ &\leq \log^{\star}(L|X_t|) + b\,,\end{aligned} \tag{A-28}$$

where $b = \log^{\star}|g^{-1}(0)|$, $\log^{\star}(x) = \log(1+x)$ and the second inequality holds for its sub-additivity. By taking the expectation

$$\mathrm{E}(\log|A'(X_t)|)_+ \leq \mathrm{E}\left(\log^{\star}(L|X_t|)\right) + \log^{\star}|g^{-1}(0)| \leq L\,\mathrm{E}|X_t| + b\,. \tag{A-29}$$

So the expectation in (A-29) is finite because the expectation of X_t is finite when $\mathrm{E}|Y_t| < \infty$, see the proof of (A-12). This proves (H1).

Assumption (H2) is required only in the context of asymptotic normality for QMLE. We remind that, if g is canonical, then $Q_t = X_t$ is the canonical parameter, and by Corollary 7, we have $A'(X_t) = \mu_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$ and $\mathrm{E}[A'(X_t)^4] = \mathrm{E}\left[\mathrm{E}(Y_t|\mathcal{F}_{t-1})^4\right] \leq \mathrm{E}\left[\mathrm{E}(Y_t^4|\mathcal{F}_{t-1})\right] = \mathrm{E}(Y_t^4) < \infty$. Then, we also have $\mathrm{E}\,|A''(X_t)| \leq |L| < \infty$, as $A'(\cdot)$ is Lipschitz, and this verifies assumption (H2). However, there are cases where the canonical link function g is not Lipschitz; for example, $g(\cdot) = \log(\cdot)$. Here the proof is immediate: $\mathrm{E}(\log|A'(X_t)|)_+ = \mathrm{E}(\log|\exp(X_t)|)_+ = \mathrm{E}|X_t| < \infty$. Moreover, $\mathrm{E}\left[A'(X_t)^4\right] = \mathrm{E}\left[A''(X_t)^4\right] \leq \mathrm{E}(Y_t^4) < \infty$.

Checking conditions (H1)-(H2) for a non-canonical link function $g(\cdot)$ clearly depends on its specific shape. We give here some relevant examples. Suppose one wants to model the expectation μ_t linearly as in (8), with a Poisson distribution coming from (1); this is done by setting $f(X_t) = \log(X_t) = \log(\mu_t)$ and $A(X_t) = X_t = \mu_t > 0$. Here, the expectations involving $A(\cdot)$ are finite, as $A'(X_t) = 1$ and $A''(X_t) = 0$. The expectations of the derivatives $f'(X_t)^4 = 1/X_t^4 \leq 1/\alpha^4$ and $f''(X_t)^4 = 1/X_t^8 \leq 1/\alpha^8$ are bounded; in fact $\mu_t > 0$, the parameters $(\alpha, \gamma, \phi, \theta) > 0$, then $X_t = \mu_t \geq \alpha$, completing the proof.

Another common model with non-canonical link function used in the literature is (9) for the Negative Binomial (10); it is derived by (1) when $d(Y_t) =$

TABLE A-1
Simulations for GLARMA(1,1); $Y_t|\mathcal{F}_{t-1} \sim Be(p_t)$, $s = 1,000$.

n		α	γ	θ	α	γ	θ	α	γ	θ
200	True	0.500	-0.400	0.800	0.500	0.400	0.200	0.500	0.400	1.200
	Est.	0.522	-0.441	0.795	0.721	0.147	0.176	0.558	0.341	1.193
	Std.Dev	0.206	0.372	0.315	1.187	1.414	0.342	0.281	0.265	0.347
	Lower	0.509	-0.464	0.776	0.647	0.059	0.154	0.541	0.324	1.172
	Upper	0.535	-0.418	0.815	0.794	0.234	0.197	0.576	0.357	1.215
	Bias	0.022	-0.041	-0.005	0.221	-0.253	-0.024	0.058	-0.059	-0.007
	KS	0.218	0.638	0.577	0.937	0.994	0.791	0.293	0.927	0.318
500	Est.	0.509	-0.432	0.791	0.604	0.274	0.184	0.517	0.381	1.189
	Std.Dev	0.124	0.219	0.187	0.762	0.911	0.207	0.168	0.171	0.219
	Lower	0.501	-0.446	0.779	0.557	0.218	0.171	0.506	0.370	1.176
	Upper	0.517	-0.418	0.803	0.651	0.331	0.197	0.527	0.391	1.203
	Bias	0.009	-0.032	-0.009	0.104	-0.126	-0.016	0.017	-0.019	-0.011
	KS	0.387	0.965	0.931	0.555	0.616	0.780	0.320	0.437	0.465
1000	Est.	0.502	-0.407	0.796	0.592	0.292	0.193	0.514	0.387	1.198
	Std.Dev	0.086	0.154	0.141	0.565	0.673	0.151	0.120	0.122	0.147
	Lower	0.496	-0.417	0.788	0.557	0.250	0.184	0.506	0.379	1.189
	Upper	0.507	-0.398	0.805	0.627	0.333	0.203	0.521	0.394	1.207
	Bias	0.002	-0.007	-0.004	0.092	-0.108	-0.007	0.014	-0.013	-0.002
	KS	0.361	0.265	0.673	0.866	0.732	0.957	0.714	0.850	0.784

$\log[\Gamma(\nu + Y_t)/(\Gamma(Y_t + 1)\Gamma(\nu))]$, $A(X_t) = -\nu \log(\nu/(\nu + \mu_t)) = \nu \log(\nu + e^{X_t}) - \nu \log(\nu)$ and $f(X_t) = \log(\mu_t/(\nu + \mu_t)) = X_t - \log(\nu + e^{X_t})$. We know that $\nu > 0$, hence $\mathrm{E}[A'(X_t)^4] = \mathrm{E}[(\nu e^{X_t}/(\nu + e^{X_t}))^4] \leq \nu^4 < \infty$ and $\mathrm{E}[A''(X_t)^4] = \mathrm{E}[(\nu^2 e^{X_t}/(\nu + e^{X_t})^2)^4] \leq \exp(\nu) < \infty$. In the same fashion, $f'(X_t)^4 = (\nu/(\nu + e^{X_t}))^4 \leq 1$ and $f''(X_t)^4 = (\nu e^{X_t}/(\nu + e^{X_t})^2)^4 \leq 1$, which have finite expectations.

Appendix B: Simulation results for finite sample properties

In this section, the numerical results concerning the finite sample properties discussed in Section 4.2 are presented. Table A-1 summarizes the estimation results for the GLARMA model when the data come from a Bernoulli distribution. Tables A-2 and A-3 show the outcome of simulations for GARMA and log-AR models performed on data generated from Geometric distribution in (10), but with Poisson distribution fitted instead (QMLE). The first row of the tables reports the true parameter values; the following two rows show the mean of the estimated parameters, obtained by averaging out the results from all simulations along with the corresponding standard error. The subsequent two rows present the lower and upper limits of the confidence interval for the estimated mean. Finally, the last two rows correspond to the bias of the mean and the p-value of the Kolmogorov-Smirnov (KS) test for normality on the standardized MLE/QLME obtained from the simulations.

TABLE A-2
Simulations QMLE of Poisson GARMA(1,1); $Y_t|\mathcal{F}_{t-1} \sim Geom(p_t)$, $s = 1,000$.

n		α	ϕ	θ	α	ϕ	θ	α	ϕ	θ
200	True	0.500	-0.400	0.800	0.500	0.400	0.200	0.500	0.400	1.200
	Est.	0.485	-0.412	0.810	0.483	0.375	0.217	0.515	0.381	1.167
	Std.Dev	0.110	0.153	0.177	0.106	0.117	0.144	0.253	0.068	0.172
	Lower	0.478	-0.421	0.799	0.476	0.367	0.209	0.499	0.377	1.156
	Upper	0.492	-0.402	0.821	0.489	0.382	0.226	0.530	0.386	1.177
	Bias	-0.015	-0.012	0.010	-0.017	-0.025	0.017	0.015	-0.019	-0.033
	KS	0.339	0.576	0.817	0.197	0.910	0.669	0.001	0.732	0.455
500	Est.	0.494	-0.406	0.806	0.492	0.392	0.204	0.497	0.392	1.192
	Std.Dev	0.065	0.102	0.115	0.067	0.077	0.091	0.200	0.051	0.127
	Lower	0.490	-0.412	0.799	0.488	0.387	0.199	0.484	0.389	1.184
	Upper	0.498	-0.400	0.813	0.496	0.396	0.210	0.509	0.395	1.199
	Bias	-0.006	-0.006	0.006	-0.008	-0.008	0.004	-0.003	-0.008	-0.008
	KS	0.418	0.566	0.640	0.851	0.963	0.285	0.000	0.375	0.015
1000	Est.	0.494	-0.401	0.800	0.493	0.395	0.203	0.504	0.395	1.187
	Std.Dev	0.048	0.071	0.080	0.046	0.054	0.066	0.169	0.041	0.108
	Lower	0.491	-0.405	0.795	0.490	0.392	0.199	0.493	0.392	1.180
	Upper	0.497	-0.396	0.805	0.496	0.398	0.207	0.514	0.397	1.194
	Bias	-0.006	-0.001	-0.000	-0.007	-0.005	0.003	0.004	-0.005	-0.013
	KS	0.272	0.370	0.549	0.984	0.936	0.988	0.000	0.198	0.050

TABLE A-3
Simulations QMLE of Poisson log-AR(1); $Y_t|\mathcal{F}_{t-1} \sim Geom(p_t)$, $s = 1,000$.

n		α	ϕ	γ	α	ϕ	γ
200	True	0.500	-0.400	0.800	0.500	0.400	0.200
	Est.	0.451	-0.411	0.858	0.553	0.385	0.155
	Std.Dev	0.219	0.130	0.266	0.274	0.110	0.237
	Lower	0.437	-0.419	0.841	0.536	0.379	0.141
	Upper	0.464	-0.402	0.874	0.571	0.392	0.170
	Bias	-0.049	-0.011	0.058	0.053	-0.015	-0.045
	KS	0.198	0.981	0.060	0.907	0.399	0.673
500	Est.	0.482	-0.401	0.820	0.528	0.395	0.177
	Std.Dev	0.133	0.077	0.165	0.176	0.065	0.144
	Lower	0.474	-0.405	0.810	0.517	0.391	0.168
	Upper	0.490	-0.396	0.830	0.539	0.399	0.186
	Bias	-0.018	-0.001	0.020	0.028	-0.005	-0.023
	KS	0.562	0.898	0.405	0.845	0.957	0.780
1000	Est.	0.488	-0.400	0.813	0.517	0.397	0.185
	Std.Dev	0.097	0.054	0.120	0.132	0.047	0.107
	Lower	0.482	-0.404	0.806	0.509	0.394	0.178
	Upper	0.494	-0.397	0.820	0.526	0.400	0.192
	Bias	-0.012	-0.000	0.013	0.017	-0.003	-0.015
	KS	0.656	0.517	0.772	0.567	0.551	0.942

Acknowledgments

We would like to thank the Editor, the Associate Editor and a Referee for valuable comments and suggestions. We also would like to thank Christian Francq, Kostas Fokianos and David Matteson for their insightful comments to an earlier version of the paper.

References

[1] Ahmad, A. and Francq, C. (2016). Poisson QMLE of count time series models. *Journal of Time Series Analysis* **37** 291-314. MR3512959

[2] Armillotta, M. (2021). Essays on discrete valued time series models, PhD thesis, University of Bologna, Italy.

[3] Benjamin, M., Rigby, R. and Stasinopoulos, D. M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association* **98** 214-223. MR1965687

[4] Box, G. E. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden Day. MR0272138

[5] Christou, V. and Fokianos, K. (2014). Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis* **35** 55–78. MR3148248

[6] Christou, V. and Fokianos, K. (2015). On count time series prediction. *Journal of Statistical Computation and Simulation* **85** 357–373. MR3270681

[7] Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics* **8** 93-115. MR0623586

[8] Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261. MR2756513

[9] Davis, R. A., Dunsmuir, W. T. M. and Streett, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika* **90** 777-790. MR2024757

[10] Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., Pipiras, V. and Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association* **116** 1533-1547. MR4309291

[11] Davis, R. A., Holan, S. H., Lund, R. and Ravishanker, N. (2016). *Handbook of Discrete-valued Time Series*. CRC Press. MR3642975

[12] Davis, R. A. and Liu, H. (2016). Theory and inference for a class of nonlinear models with application to time series of counts. *Statistica Sinica* **26** 1673-1707. MR3586234

[13] Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM* **41** 45-76. MR1669737

[14] Douc, R., Doukhan, P. and Moulines, E. (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications* **123** 2620 - 2647. MR3054539

[15] Douc, R., Fokianos, K. and Moulines, E. (2017). Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electronic Journal of Statistics* **11** 2707–2740. MR3679907
[16] Doukhan, P., Fokianos, K. and Tjøstheim, D. (2012). On weak dependence conditions for Poisson autoregressions. *Statistics & Probability Letters* **82** 942–948. MR2910041
[17] Dunsmuir, W. and Scott, D. (2015). The GLARMA package for observation-driven time series regression of counts. *Journal of Statistical Software* **67** 1-36.
[18] Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH process. *Journal of Time Series Analysis* **27** 923–942. MR2328548
[19] Fokianos, K. (2022). Multivariate Count Time Series Modelling. *To appear in Econometrics and Statistics*.
[20] Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association* **104** 1430–1439. MR2596998
[21] Fokianos, K., Støve, B., Tjøstheim, D. and Doukhan, P. (2020). Multivariate count autoregression. *Bernoulli* **26** 471–499. MR4036041
[22] Fokianos, K. and Tjøstheim, D. (2011). Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102** 563–578. MR2755016
[23] Fokianos, K. and Truquet, L. (2019). On categorical time series models with covariates. *Stochastic processes and their applications* **129** 3446–3462. MR3985569
[24] Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69** 243–268. MR2325275
[25] Gorgi, P. (2020). Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations. *Journal of the Royal Statistical Society: Series B*. MR4176345
[26] Inouye, D. I., Yang, E., Allen, G. I. and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics* **9** 1–25. MR3648601
[27] Li, W. K. (1994). Time series models based on generalized linear models: some further results. *Biometrics* **50** 506-511.
[28] Liboschik, T., Fokianos, K. and Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software* **82** 1–51.
[29] Matteson, D. S., Woodard, D. B. and Henderson, S. G. (2011). Stationarity of generalized autoregressive moving average models. *Electronic Journal of Statistics* **5** 800-828. MR2824817
[30] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall. MR3223057
[31] Neumann, M. H. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17** 1268-1284. MR2854772
[32] Pan, W. (2001). Akaike's information criterion in generalized estimating

equations. *Biometrics* **57** 120–125. MR1833297

[33] Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys* **1** 20-71. MR2095565

[34] Rydberg, T. H. and Shephard, N. (2003). Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics* **1** 2-25.

[35] Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data. *Annals of Applied Statistics* **4** 943-961. MR2758428

[36] Shephard, N. (1995). Generalized linear autoregressions. Unpublished paper.

[37] Slutsky, E. (1937). The summation of random causes as the source of cyclic processes. *Econometrica: Journal of the Econometric Society* 105–146.

[38] Startz, R. (2008). Binomial autoregressive moving average models with an application to U.S. recessions. *Journal of Business & Economic Statistics* **26** 1-8. MR2422056

[39] Thorisson, H. (1995). Coupling methods in probability theory. *Scandinavian Journal of Statistics* **22** 159-182. MR1339749

[40] Tweedie, R. L. (1988). Invariant measures for Markov chains with no irreducibility assumptions. *Journal of Applied Probability* **25** 275–285. MR0974587

[41] Walker, G. T. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A* **131** 518–532.

[42] Yule, G. U. (1927). On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A* **226** 267–298.

[43] Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* **44** 1019–1031. MR0980997

[44] Zheng, T., Xiao, H. and Chen, R. (2015). Generalized ARMA models with martingale difference errors. *Journal of Econometrics* **189** 492 - 506. MR3414917