Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo 33

Settore Concorsuale: 13/D1 - STATISTICA

Settore Scientifico Disciplinare: SECS-S/01 - STATISTICA

ESSAYS ON DISCRETE VALUED TIME SERIES MODELS

Presentata da: Mirko Armillotta

Coordinatore Dottorato

Monica Chiogna

Supervisore

Alessandra Luati

Co-Supervisore

Monia Lupparelli

Esame finale anno 2021

Abstract

Statistical inference for discrete-valued time series has not been developed as systematically as traditional methods for time series generated by continuous random variables. This Ph.D. dissertation deals with time series models for discrete-valued processes. In particular, Chapter 2 is devoted to a comprehensive overview of the literature about observation-driven models for discrete-valued time series. Derivation of stochastic properties for these models is presented. For the inference, general properties of the quasi maximum likelihood estimator (QMLE) are discussed, followed by an illustrative application.

In Chapter 3, a general class of observation-driven time series models for discrete-valued processes is introduced. Stationarity and ergodicity are derived under easy-to-check conditions, which can be directly applied to all the models encompassed in the framework. Consistency and asymptotic normality of the QMLE are established, with the focus on the exponential family. Finite sample properties of the estimators are investigated through a Monte Carlo study and illustrative examples are provided. The framework introduced in the paper provides a self-contained background that relates different models developed in the literature as well as novel specifications and makes them fully applicable in practice.

Discrete responses are commonly encountered in real applications and are strongly connected to network data. The specification of suitable network autoregressive models for count time series is an important aspect which is not covered by the existing literature. In Chapter 4, we consider network autoregressive models for count data with a known neighborhood structure. The main methodological contribution is the development of conditions that guarantee stability and valid statistical inference. We consider both cases of fixed and increasing network dimension and we show that quasi-likelihood inference provides consistent and asymptotically normally distributed estimators. The work is complemented by simulation results and a data example.

Contents

Chapter 1

Introduction

In recent years the availability of discrete data coming from several sources have motivated the outset of a wide literature on models for count time series processes. A growing attention to high dimensional data sets involving dynamic binary and count data has been object of interest, in different contexts. For example, the number of clicks or amount of intra-day stock transactions (Davis and Liu, 2016; Ahmad and Francq, 2016). Besides that, time series analysis for integer valued random variables has not been developed as the continuous counterpart, which, instead, has a long tradition in time series analysis. The peculiar discrete nature of the process requires an ad hoc treatment for the development of the asymptotic theory of the estimators. The same is true for the establishment of probabilistic properties, strict stationarity and ergodicity, of integer valued processes. Other central aspects are related to the establishment of a comprehensive inferential theory as well as a robust model selection procedure between several candidate models, so as to make these model fully applicable in real studies. Moreover, usual concepts of the continuous time series analysis, such as auto-covariance or the Wold representation, need and adapted definition or can be meaningless.

Time series models for discrete data can be divided in two families of models: observation driven models and parameter driven models. This thesis deals focuses on the observation driven models (Cox, 1981); which are described by a discrete time series process and a latent process, the latter is defined as pure deterministic function of the former's past history. In the parameter driven models, instead, the latent process depends on unknown parameters and it is treated as stochastic.

The rest of the PhD dissertation is organized as follow. In Chapter 2 a survey of the most famous time series models for integer valued processes is presented. Chapter 3 introduces a general modelling framework on observation driven models for discrete data, as an original scientific article. Then, Chapter 4 regards a new contribution on network autoregression models for Poisson processes. Finally, Chapter 5 hosts some concluding remark on future directions of research.

More precisely, Chapter 2 is devoted to a comprehensive overview of a wide class of observation driven models for discrete valued time series, with special focus on count and binary data. In particular, technical and modelling properties are discussed for ARMA-like time series models for integer valued processes Benjamin et al. (2003); Davis et al. (2003); Startz (2008). The use of these ARMA-like models is illustrated through the analysis of the daily number of deaths for COVID-19 in Italy from March to December 2020.

The analysis is performed under the assumption both of a Poisson and of a Negative Binomial distribution for the data generating process. Finally, model comparison is carried out by using penalized likelihood criteria.

Recent developments on binary and count times series models, involving several approaches and different specifications for a wide range of models established a fragmentary literature. There would be a benefit from the specification of a unified framework able to encompass most of the models available in the literature. This will enable to study relations among models and to derive an unified approach for the derivation of stochastic properties holding across all the models. Some authors have provided a remarkable formulation of a general framework for observation driven models, with specific focus on discrete data, see Douc et al. (2013). However, this theoretical formulation might be not effective when the aim is to implement models in real practices. More precisely, the ergodicity conditions established by Douc et al. (2013) are hard to verify in practice, and they vary for each model and every different distribution.

Then, in Chapter 3, we introduce a general modelling framework aiming to provide a unified specification for a general class of integer valued time series. From this general framework we point out some special cases of particular interest, which are new models not directly presented in the literature yet. Then, we analyze the relationships among different models belonging to the framework. Furthermore, stochastic properties which hold simultaneously for the entire class of models are derived (strict stationarity and ergodicity). For some of them, stability conditions have not been set in the literature yet. Finally, a quasi-maximum likelihood (QMLE) inference is provided with the asymptotic properties of the estimator. These results make all the models encompassed in the framework fully applicable in practice.

Further sources of information gaining remarkable importance are constituted by network data, which are considered of essential importance for many topic of research (social network, epidemics, etc..). In particular, quantifying the impact of a network structure-like dependence on a time series process raises critical interest. Discrete variables are usually detected in the practice of network studies. For example, several information of interest in social network analysis have an integer nature. Then, binary and count processes are substantially related with network data. As far as we know, at the present time, no such models exist for non-continuous responses, even though a flourishing literature for their continuous counterparts has been set, see Zhu et al. (2017). This is an a crucial open space in the present literature.

The main aim of Chapter 4 is exactly to fill this lack in the literature by specifying a linear and a log-linear version of the Poisson network autoregression (PNAR) for count processes. We even derive minimal stability properties of such models. Moreover, in this field two types of asymptotic inference are possible: with increasing time sample size and fixed network dimension and with both time and network dimension increasing together. The QMLE is established for the PNAR models under both types of asymptotics. A further aspect of interest is that all the network time series models presented so far are defined under the i.i.d. assumption of the error terms. This might be not realistic in many empirical applications. We overtake this limit by employing the concept of α-mixing (see Doukhan (1994)) which is a measure of *asymptotic independence* over a timespan, allowing to relax the i.i.d. assumption. Then, a complex and flexible dependence structure among variables is specified, among time and among the network, and this is effected by defining a copula construction for modelling the dependence between variables.

Bibliography

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

Benjamin, M., R. Rigby, and D. Stasinopoulos (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association 98*(461), 214–223.

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

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

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

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

Doukhan, P. (1994). *Mixing*, Volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

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

Zhu, X., R. Pan, G. Li, Y. Liu, and H. Wang (2017). Network vector autoregression. *The Annals of Statistics 45*, 1096–1123.

Chapter 2

An overview of ARMA-like models for count and binary data

Mirko Armillotta[1], Alessandra Luati[1] and Monia Lupparelli[2]

[1] *Department of Statistical Sciences, University of Bologna, 41 st. Belle Arti, 40126, Bologna, Italy.*
Email: mirko.armillotta2@unibo.it, alessandra.luati@unibo.it

[2] *Department of Statistics, Computer Science, Applications, University of Florence, 59 ave. Morgagni, 50134, Florence, Italy.*
Email: monia.lupparelli@unifi.it

Abstract

A comprehensive overview of the literature on models for discrete valued time series is provided, with a special focus on count and binary data. ARMA-like models such as the BARMA, GARMA, M-GARMA, GLARMA and log-linear Poisson are illustrated in detail and critically compared. Methods for deriving the stochastic properties of specific models are delineated and likelihood-based inference is discussed. The review is concluded with an empirical application, concerned with the analysis of the daily number of deaths for COVID-19 in Italy, under the assumption both of a Poisson and a negative binomial distribution for the data generating process.

2.1 Introduction

Traditionally, time series modelling has been mostly applied to data that are continuously valued. From the early specifications of Yule (1927) and Walker (1931), to the formalisation by Box and Jenkins (1970, 1976), autoregressive (AR) and moving average (MA) models have been regularly applied in many fields, from finance to energy and neural networks, see for example Ho et al. (2002), Wang et al. (2012) and Sen et al. (2016). Non-linear models, such as the generalized autoregressive conditional heteroskedastic models by (Engle (1982), Bollerslev (1986)) or the threshold and smooth transition models (Tong and Lim (1980), Teräsvirta (1994)), up to the class of score driven models (Creal et al. (2013), Harvey (2013)), are essentially grounded on autoregressive dynamics. Though often employed regardless of the discrete nature of the data generating process, continuous models do not adequately describe the dynamic trend of count or binary data. Notable examples where *ad hoc* models for discrete data are required include the number of clicks on a website and the daily counts of people infected with a rare disease or, as far as binary

data are concerned, the presence or absence of an edge in a random network system and the success or failure of an industrial process.

Despite some relevant instances that we aim to discuss in this chapter, ARMA models for discrete valued time series have not enjoyed the same popularity of linear models for continuous time series. One of the reasons certainly lies in the fact that linear processes are related to second order stationarity, which fully characterizes Gaussian time series, while for discrete or count data, the concept of autocovariance needs to be adapted Startz (2008). Moreover, the Wold representation, which allows every covariance-stationary time series to be written as the sum of two time series, one deterministic and one stochastic, has no direct interpretation Davis et al. (2016) in the integer-valued case. As a matter of fact, modelling discrete valued time series entails challenging aspects which are directly related to the nature of the generating random process.

In recent years, the interest in the analysis of discrete dynamic data has been considerably increasing. An useful classification of time series models in two main families is due to Cox (1981), who distinguished observation driven models (see Zeger and Liang (1986)) and parameter driven models (Zeger (1988)). In the parameter driven models two different time series processes are object of inference: the process of the observed data, say $\{Y_t\}_{t\in Z}$, and an unobservable latent time series $\{\mu_t\}_{t\in Z}$ which presents a dynamic formulation and carry an error term $\{e_t\}_{t\in\mathbb{Z}}$. The observation driven models, instead, are fully described by the time series of the observed process $\{Y_t\}_{t\in Z}$, since here the latent process $\{\mu_t\}_{t\in Z}$ is simply defined as a deterministic function of the past history of Y_t.

An early contribution to the development of integer valued time series is constituted by Integer Autoregressive models (INAR) Al-Osh and Alzaid (1987); Alzaid and Al-Osh (1990), that is categorized as an observation driven model. Some other examples of observation-driven models for count time series include the works by Davis et al. (2003), Benjamin et al. (2003) and Ferland et al. (2006), among others. With the focus on the dynamic trend of count data, recent contributions can be envisaged in the works of Rydberg and Shephard (2003), Kauppi and Saikkonen (2008), Davis and Liu (2016), Ahmad and Francq (2016) and Clark et al. (2018) and Gorgi (2020).

The aim of this chapter is to provide a comprehensive overview of the literature on observation driven models for discrete valued time series, with a special focus on count and binary data. In particular, stochastic properties and estimation are discussed for notable ARMA-like models, such as BARMA Li (1994), GARMA Benjamin et al. (2003), GLARMA Davis et al. (2003), M-GARMA Zheng et al. (2015) and log-linear Poisson Fokianos et al. (2009) models. These models are generally referred to ARMA-like models as they are designed to account for the direction and the magnitude of three relevant effects in the analysis of temporal data. More precisely, ARMA-like models may include an autoregressive-like effect, a moving average type effect and the dependence with respect to the past predictions of the random process. The specification for these effects eventually depends on suitable link functions which are selected according to the probabilistic assumptions for the data generating process.

The stochastic properties of discrete ARMA models can be derived following two different methods based on the Markov chain theory and the perturbation approach, among others. The perturbation approach developed by Fokianos et al. (2009) is based on the analysis of a modified version of the discrete process, which allows one to derive properties of the original processes. An alternative method, based on Markov chain theory without irreducibility assumptions, has been considered by Matteson et al. (2011) and Douc et al. (2013). This approach leads to obtaining probabilistic properties of the discrete variable by defining the latent process as a Markov chain of order one. To illustrate these methods, an example for the GARMA model is given, taken from Matteson et al. (2011). An application to log-linear Poisson autoregression provided by Douc et al. (2013) is reported, as well.

As far as inference is concerned, the properties of the maximum likelihood estimator (MLE) and Quasi MLE (QMLE) have been widely studied for discrete-valued models; see Douc et al. (2013), Davis and Liu (2016) and Ahmad and Francq (2016), among others. Specifically, the use of the generalized linear model (GLM) of McCullagh and Nelder (1989) for dynamic discrete data provides a natural extension of continuous-valued time series to integer-valued processes. Then, theory for likelihood inference can be acquired directly from the GLM framework as well as

principles for hypothesis testing and model diagnostics. For the case of misspecified models, results related to quasi likelihood inference are also illustrated, together with the conditions required for strong consistent and asymptotically normal QMLE, based on the work of Douc et al. (2013) and Douc et al. (2017). Clearly, the exact likelihood inference and the asymptotic properties of the MLE are obtained as a special case.

To conclude the review, an application of the ARMA-like models is illustrated through the analysis of the recent time series related to the daily number of deaths for COVID-19 in Italy from March to December 2020. The analysis is performed under the assumption of a Poisson and a negative binomial distribution for the data generating process. Model comparison is carried out by using penalized likelihood criteria.

2.2 General overview

Let us consider a stochastic process $\{Y_t\}_{t\in\mathbb{N}}$, the information set of past observations of the process $\mathcal{F}_{t-1} = \sigma\{(\mathbf{X}_{s+1}, Y_s),\ s \leq t-1\}$ up to the time $t-1$ and a vector of covariates $\mathbf{X}_t$ up to time t, where $\sigma\{X\}$ refers to the sigma-field generated by the random variable X, and it is defined as the smallest sigma-field with respect to which it is measurable. For the definition of sigma-field see (Billingsley, 1995, p. 19-20). The corresponding realizations are denoted with the lower-case counterparts, y_t and $\mathbf{x}_t$, respectively. The focus, throughout the chapter, is on the case when $\{Y_t\}_{t\in\mathbb{N}}$ is discrete-valued. Suppose that the distribution of the process lies in the general class of one-parameter exponential family:

$$q(Y_t\,|\mathcal{F}_{t-1}\,) = \exp\{Y_t\, f(\eta_t) - A(\eta_t) + d(Y_t)\}\ , \tag{2.1}$$

where the conditional expected value is defined as

$$\mu_t = \mathrm{E}(Y_t\,|\mathcal{F}_{t-1}\,) = A^{'}(\eta_t)$$

and $\eta_t = g(\mu_t)$ with $g(\cdot)$ a twice-differentiable, one-to-one monotonic function, which is called link function, see McCullagh and Nelder (1989).

In equation (2.1) it is assumed that the dynamics of the density (or mass) function $q(Y_t|\mathcal{F}_{t-1})$ are captured by the parameter μ_t, or equivalently η_t, called linear predictor. The function $A(\cdot)$ (log-partition) and $d(\cdot)$ are specific functions which define the particular distribution of interest. In the framework of the exponential family of McCullagh and Nelder (1989), $f(\eta_t)$ is the canonical parameter. The mapping $f(\cdot)$ is a twice-differentiable bijective function, chosen accordingly to the model of interest. The conditional variance is

$$\sigma_t^2 = \mathrm{V}(Y_t\,|\mathcal{F}_{t-1}\,) = A^{''}(\eta_t) = v(\mu_t)\,.$$

Example 1. *In equation* (2.1)*, the Poisson distribution is obtained by setting* $f(\eta_t) = \eta_t$*,* $\eta_t = g(\mu_t) = \log(\mu_t)$*,* $A(\eta_t) = \exp(\eta_t) = \mu_t$ *and* $d(Y_t) = \log(1/Y_t!)$*. The conditional expectation is then* $\mathrm{E}(Y_t|\mathcal{F}_{t-1}) = \mathrm{V}(Y_t|\mathcal{F}_{t-1}) = \exp(\eta) = \mu_t$*.*

Clearly, since for the Poisson distribution the canonical parameter is $\eta_t = \log(\mu)$, see McCullagh and Nelder (1989), one has $f(\eta_t) = \eta_t$.

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

It can be convenient to consider the following dynamic representation for the time varying conditional mean,

$$g(\mu_t) = \eta_t = \mathbf{x}_t^T\beta + z_t\,, \tag{2.2}$$

$$z_t = \sum_{j=1}^{p} \phi_j \left[h(Y_{t-j}) - \mathbf{x}_{t-j}^T \beta \right] + \sum_{j=1}^{k} \gamma_j (z_{t-j} + \epsilon_{t-j}) + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \,, \tag{2.3}$$

where p, k adn q are integers representing the maximum lag order of their respective additive terms, and ϵ_t, generally called *prediction error*, is defined in the following way:

$$\epsilon_t = \frac{h(Y_t) - \bar{g}(\mu_t)}{\nu_t} \tag{2.4}$$

and ν_t is some scaling sequence, for example:

- $\nu_t = \sigma_t$, Pearson residuals
- $\nu_t = \sigma_t^2$, Score-type residuals
- $\nu_t = 1$, No scaling
- $\nu_t = \mathrm{V}[h(y_t)\,|\mathcal{F}_{t-1}]$

where $\mathrm{V}[h(y_t)\,|\mathcal{F}_{t-1}]$ is the variance of the function $h(Y_t)$, conditional to the past information $\mathcal{F}_{t-1}$.

Furthermore, the function $h(Y_t)$ is called "data-link function" because it is applied to the observation process Y_t whereas $\bar{g}(\mu_t)$ is said "mean-link function" because 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; their shape depends on the specific model (2.2)-(2.3) and the distribution of interest in equation (2.1). Note that the terminology link function is generally referred to the specification of a function $g(\cdot)$ for modelling the dependence between a transformation η_t of the conditional expected value μ_t and a linear predictor including information related to past values z_t or to a covariate set $\mathbf{x}_t$. The same terminology is here adopted for the specification of functions $h(\cdot)$ and $\bar{g}(\cdot)$ since, in some instances belonging to the exponential family distribution, convenient choices for these functions correspond to the canonical link function. Nevertheless, $h(\cdot)$ and $\bar{g}(\cdot)$ might be different from $g(\cdot)$, so that the model (2.2)-(2.3) is able to encompass a wide range of existing models developed in the literature, as its special cases. Some examples are presented in the next section.

Despite the fact that it is not constrained to assume a specific formulation, in general, it is useful to choose the mean-link function as follows:

$$\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)\,|\mathcal{F}_{t-1}]\,, \tag{2.5}$$

in order to obtain $\epsilon_t \sim MDS$ (Martingale Difference Sequence), i.e. the difference $\mathrm{E}[h(Y_t) - \bar{g}(\mu_t)|\mathcal{F}_{t-1}] = 0$. In fact, a MDS process has conditional expectation $\mathrm{E}[\epsilon_t\,|\mathcal{F}_{t-1}] = 0$ and unconditional expectation $\mathrm{E}(\epsilon_t) = 0$. Moreover it is uncorrelated, i.e. $\mathrm{E}(\epsilon_t \epsilon_{t-s}) = 0$, with $s \neq 0$. This is a really useful construct in probability theory because it does not require the usual assumption of independence of the errors. Furthermore, most limit theorems that hold for an independent sequence will also hold for a MDS.

Moreover, if $\nu_t = \sqrt{\mathrm{V}[h(Y_t)\,|\mathcal{F}_{t-1}]}$, then the residuals in equation (2.4) form a white noise (WN) sequence, with unit variance. In practical situations, an explicit formula for the conditional moments $\mathrm{E}\left[h(Y_t)|\mathcal{F}_{t-1}\right]$ and $\mathrm{V}\left[h(Y_t)|\mathcal{F}_{t-1}\right]$ is not always available. In this cases, it seems reasonable to use an approximation constructed from their Taylor expansions; for example, the second order expansions are: $\bar{g}(\mu_t) = \mathrm{E}\left[h(Y_t)|\mathcal{F}_{t-1}\right] \approx h(\mu_t) + \frac{1}{2}h''(\mu_t)\,\sigma_t^2$, $\mathrm{V}\left[h(Y_t)|\mathcal{F}_{t-1}\right] = \mathrm{E}\left[h(Y_t)^2|\mathcal{F}_{t-1}\right] - \mathrm{E}\left[h(Y_t)|\mathcal{F}_{t-1}\right]^2 \approx m(\mu_t) + \frac{1}{2}m''(\mu_t)\,\sigma_t^2 - \bar{g}(\mu_t)^2$, where $m(\cdot) = h(\cdot)^2$.

Note that the process $\{Y_t\}_{t\in\mathbb{N}}$ is observed whereas $\{\mu_t\}_{t\in\mathbb{N}}$ is not. However, it can be shown by backward substitutions in (2.2)-(2.3), that the process $\{\mu_t\}_{t\in\mathbb{N}}$ is a deterministic function of the past $\mathcal{F}_{t-1}$. This is the reason why equations (2.2)-(2.3) belong to the class of "observation-driven models", see Cox (1981).

The parameters ϕ, θ and γ in equation (2.3) model the direction and the magnitude of three relevant effects in the analysis of temporal data. Firstly, the autoregressive-like effect which represents the dependence on the

past observations; then, the effect of the moving average part is considered for modelling the dependence between prediction error terms over time; finally, the effect of the past memory dependence accounts for the dependence with respect to the past prediction rather than on the past observations. The latter can be seen as the dependence of the process from its whole past (since μ_t depends on all the past observations $Y_{t-1}, Y_{t-2}, \dots$). In principle, any effect can be specified in the model through different link functions. Typically, these functions are tailored to the nature of the data generating process.

2.3 Some relevant models

This section describes the most relevant models developed in the literature of ARMA-like time series for binary and count observations generated from probability distributions mainly belonging to the exponential family.

2.3.1 GARMA

A well-known specification for discrete-valued time series is the generalized Autoregressive Moving Average model, GARMA, Benjamin et al. (2003). Here, the distribution of the process is defined to be the one-parameter exponential family (2.1). From equation (2.2)-(2.3) the GARMA model is obtained when $k = 0$, by setting $g \equiv \bar{g} \equiv h$ and $\nu_t = 1$, so that, the three link functions are equivalent and no scaling is applied:

$$\eta_t = \mathbf{x}_t^T \beta + \sum_{j=1}^{p} \phi_j \left[g(Y_{t-j}) - \mathbf{x}_{t-j}^T \beta \right] + \sum_{j=1}^{q} \theta_j \left[g(Y_{t-j}) - \eta_{t-j} \right]. \tag{2.6}$$

The model includes the autoregressive and the moving average effects by using the same link function g. The dependence with the past memory is not considered directly by a specific factor. This means that model (2.6) would be employed when the immediate past values of the observed process $Y_{t-j}, j = 1, \dots, \max(p, q)$ may be considered influential. In general, ϵ_t is not a martingale difference sequence then the mean-link function $\bar{g}$ here does not follows (2.5), instead, it is just set to be equivalent to g. However, there still is a special case in which $\epsilon_t \sim MDS$, such as $g \equiv h : \textit{identity}$ (see the M-GARMA model below).

Although this model is suitably applicable in practice to every distribution encompassed in (2.1), it has been mainly used for count data following a Negative Binomial (NB) distribution like equation (12) in Benjamin et al. (2003).

2.3.2 M-GARMA

A suitable extension of the GARMA model in (2.6) has recently been introduced by Zheng et al. (2015); it allows the residuals ϵ_t to be a martingale difference sequence (MDS), for this reason it has been called martinagalised GARMA (M-GARMA). It is obtained from (2.2)-(2.3) for $k = 0$, $g(\mu_t) = \mathrm{E}[h(y_t) \,|\mathcal{F}_{t-1}] = \bar{g}(\mu_t)$ and $\nu_t = 1$:

$$\bar{g}(\mu_t) = \mathbf{x}_t^T \beta + \sum_{j=1}^{p} \phi_j \left[h(Y_{t-j}) - \mathbf{x}_{t-j}^T \beta \right] + \sum_{j=1}^{q} \theta_j \left[h(Y_{t-j}) - \bar{g}(\mu_{t-j}) \right]. \tag{2.7}$$

For its particular construction, in this model the crucial choice is on the data-link function h which would entirely determine the mean-link function. The usefulness of this model is on the possibility to write $h(Y_t)$ as a standard ARMA model simply by adding $h(Y_t) - \bar{g}(\mu_t)$ in both sides of (2.7) and rearranging the covariates:

$$h(Y_t) = \mathbf{x}_t^T \alpha + \sum_{j=1}^{p} \phi_j h(Y_{t-j}) + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j},$$

where $\alpha = \left(1 - \sum_{j=1}^{p} \phi B^j\right) \beta$ and B is the lag operator, such as $B^j \mathbf{x}_t = \mathbf{x}_{t-j}$. Note that when $\bar{g}(\mu_t) = \mathrm{E}[h(y_t)\,|\mathcal{F}_{t-1}] = h(\mu_t)$, a GARMA model with the linear predictor $\eta_t = \mathrm{E}[h(y_t)\,|\mathcal{F}_{t-1}]$ is obtained. Also, the use of the first-order Taylor approximation for $\bar{g}(\cdot)$ around μ_t provides

$$\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)\,|\mathcal{F}_{t-1}] \approx h(\mu_t)\,.$$

Then, the standard GARMA model has been found as a particular case of the M-GARMA model when linear approximation of $\bar{g}$ is used. This leads to consider the application of model (2.7), instead of the usual GARMA model (2.6), in all the cases when the expression $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)\,|\mathcal{F}_{t-1}]$ has a closed-form. This happens only under certain distributions, (such as Lognormal, Gamma and Beta, among others) and suitable choices the data-link function $h(\cdot)$. The interested reader can find an exhaustive treatment of such particular cases under (Zheng et al., 2015, Tab. 1).

2.3.3 GLARMA

A promising class has been developed by Rydberg and Shephard (2003) and Davis et al. (2003) under the name of generalized Linear Autoregressive Moving Average (GLARMA) models; here, again, the distribution belongs to the exponential family (2.1). GLARMA models can be written based on equations (2.2)-(2.3) by setting $p = 0$ and $h : \mathit{identity}$:

$$\begin{aligned} \eta_t &= \mathbf{x}_t^T \beta + z_t, \\ z_t &= \sum_{j=1}^{k} \gamma_j (z_{t-j} + \epsilon_{t-j}) + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \\ \epsilon_t &= \frac{Y_t - \mu_t}{\nu_t}. \end{aligned} \tag{2.8}$$

In this models, the error component and the past lag of the latent process are considered. However, the effect of past lags of the discrete process Y_t are not directly specified in the model. Notice that this model is equivalent to an ARMA model on the linear predictor (minus the constants and covariates):

$$\eta_t - \mathbf{x}_t^T \beta = z_t = \sum_{j=1}^{k} \gamma_j z_{t-j} + \sum_{j=1}^{\tilde{q}} \tau_j \epsilon_{t-j}\,,$$

where $\tilde{q} = \max(k, q)$ and $\tau_j = \gamma_j + \theta_j$. Or alternatively, in terms of η_t, we have

$$\eta_t = \mathbf{x}_t^T \alpha + \sum_{j=1}^{k} \gamma_j \eta_{t-j} + \sum_{j=1}^{\tilde{q}} \tau_j \epsilon_{t-j}\,, \tag{2.9}$$

where $\alpha = \left(1 - \sum_{j=1}^{k} \gamma_j B^j\right) \beta$.

2.3.4 Log-linear Poisson autoregression

Poisson autoregression, henceforth Pois AR, introduced by Fokianos et al. (2009), is obtained when (2.1) is $Pois(\mu_t)$, with $f(\eta_t) = \log(\eta_t)$, and in equation (2.2)-(2.3), one has $q = 0$ and $g \equiv h : \mathit{identity}$:

$$\mu_t = \mathbf{x}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \mu_{t-j} + \sum_{j=1}^{p} \phi_j Y_{t-j}\,. \tag{2.10}$$

Obviously, the parameters in equation (2.10) are constrained in the positive real line. A variant of (2.10) is the log-linear Poisson autoregression, henceforth Pois log-AR, Fokianos and Tjøstheim (2011) which is obtained when $q = 0$, $f(\eta_t) = \eta_t$, $g(\mu_t) = \log(\mu_t)$ and $h(Y_t) = \log(Y_t + 1)$:

$$\log(\mu_t) = \mathbf{x}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \log(\mu_{t-j}) + \sum_{j=1}^{p} \phi_j \log(Y_{t-j} + 1)\,. \tag{2.11}$$

The models (2.10) and (2.11) consider lagged effects for the discrete variable and the mean process explicitly and do not include an error component. However, note that, for Poisson data, the GARMA model (2.6) with identity or log links can be considered as a constrained Poisson autoregression where $\gamma_j = -\theta_j$ and ϕ_j is replaced by $\phi_j + \theta_j$, in equations (2.10) or (2.11). So that the Poisson autoregression model can be rewritten in ARMA form.

The model in (2.11) could be used also for Negative Binomial data, by rewriting the distribution in terms of the expected value parameter μ_t, see Christou and Fokianos (2014):

$$q\left(Y_t|\mathcal{F}_{t-1}\right) = \frac{\Gamma(\nu + Y_t)}{\Gamma(Y_t + 1)\Gamma(\nu)} \left(\frac{\nu}{\nu + \mu_t}\right)^{v} \left(\frac{\mu_t}{\nu + \mu_t}\right)^{Y_t} \tag{2.12}$$

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 (2.12) with model (2.11) is obtained from the distribution (2.1), by setting the non-canonical link $g(\mu_t) = \log(\mu_t)$ and $f(\eta_t) = \eta_t - \log(\nu + e^{\eta_t})$, with $A(\eta_t) = -\nu \log\left(\frac{\nu}{\nu+e^{\eta_t}}\right)$ and $d(Y_t) = \log \frac{\Gamma(\nu+Y_t)}{\Gamma(Y_t+1)\Gamma(\nu)}$.

2.3.5 BARMA

In case of dynamic binary data, a relevant model is the Binomial ARMA (BARMA) model (Li (1994), Startz (2008)) which is obtained when (2.1) is $Bin(a, \mu_t)$, where the number of trials a is known and the probability parameter is $p_t = \mu_t/a$. By setting $k = 0$, $h : identity$ and $\nu_t = 1$ in (2.2)-(2.3), we have

$$\eta_t = \mathbf{x}_t^T \beta + \sum_{j=1}^{p} \phi_j \left[Y_{t-j} - \mathbf{x}_{t-j}^T \beta\right] + \sum_{j=1}^{q} \theta_j \left[Y_{t-j} - \mu_{t-j}\right].$$

Note that, when $h : identity$, the mean-link function in (2.5) automatically reduces to $\mathrm{E}(Y_t\,|\mathcal{F}_{t-1}) = \mu_t$. Instead, the link function g can be any suitable function, typically logit or probit. This model is thought for Binomial distribution in (2.1). BARMA model includes the autoregressive effect and the moving average part. The model could be also generalized to consider the dependence with respect to the long memory term with a suitable link function.

Models for binary time series have not enjoyed the same developments as models for count data. However, enhancements in this direction could provide useful insights in several fields. The generalization for the non-binary case could be also interesting for the analysis of temporal categorical data. To the best of our knowledge this part of the literature seems to be barely explored; see Fokianos et al. (2003) and Moysiadis and Fokianos (2014) for an introduction to these models.

2.4 Weak stationarity

We now pass to examine stationarity and ergodicity for some of the models highlighted in the previous section. Specifically, we consider weak stationarity conditions for GARMA, M-GARMA and GLARMA models, in this section. For the BARMA model, no direct results on weak stationarity are available in the literature so far. However, strong stationarity is proved for BARMA, see Moysiadis and Fokianos (2014), that we shall consider in Section 2.5 along with the Poisson autoregression, derived by Fokianos et al. (2009) and Fokianos and Tjøstheim (2011).

2.4.1 GARMA

For the GARMA model in (2.6) for $g \equiv h : identity$, one has $\epsilon_t = Y_t - \mu_t$, with zero conditional and unconditional mean value. Moreover the process ϵ_t is uncorrelated. The observation process can now be expressed in the form

$$Y_t = \mu_t + \epsilon_t. \tag{2.13}$$

By setting $w_t = Y_t - \mathbf{x}_t^T \beta$ and by replacing the expression of (2.6) in (2.13), a standard ARMA model is obtained:

$$w_t = \sum_{j=1}^{p} \phi_j \, w_{t-j} + \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j} + \epsilon_t. \tag{2.14}$$

Of course (2.14) can be easily rearranged via polynomial notation in:

$$w_t = \Psi(B) \, \epsilon_t$$

where $\Psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots = \Phi(B)^{-1} \, \Theta(B)$, $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, $\Theta(B) = 1 + \theta_1 B - \cdots - \theta_p B^q$ and B is the lag operator; provided that $\Phi(B)$ is invertible. Indeed, look that $\mathrm{E}(w_t) = \Psi(B)$, $\mathrm{E}(\epsilon_t) = 0$ and then $\mathrm{E}(Y_t) = \beta$ in the case where $\mathbf{x}_t^T \beta = \beta$. The autocovariance does not depend on time t because of the uncorrelated ϵ_t. Concerning the variance the situation is more complex:

$$\begin{aligned}
\mathrm{V}(Y_t) &= \mathrm{V}(\mathbf{x}^T \beta + w_t) \\
&= \mathrm{V}(w_t) = \mathrm{E}(\epsilon_t^2) \\
&= \mathrm{E}\left[\Psi(B) \, \epsilon_t \, \Psi(B) \, \epsilon_t\right] \\
&= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i \, \psi_j \mathrm{E}(\epsilon_{t-i} \, \epsilon_{t-j}) \\
&= \sum_{i=0}^{\infty} \psi_i^2 \mathrm{E}(\epsilon_{t-i}^2) \\
&= \varphi \mathrm{E}\left[\Psi^{(2)}(B) \, \upsilon(\mu_t)\right],
\end{aligned} \tag{2.15}$$

where $1 + \psi_1^2 B + \psi_2^2 B^2 + \cdots = \Psi^{(2)}(B)$. Expression (2.15) is obtained remembering that $\mathrm{E}(\epsilon_t^2) = \mathrm{V}(\epsilon_t) = \mathrm{E}\left[\mathrm{E}(\epsilon_t^2 \mid \mathcal{F}_{t-1})\right] = \mathrm{E}\left[\upsilon(\mu_t)\right]$. The expression of the unconditional variance for the mean can be found as follows:$\mathrm{V}(Y_t) = \mathrm{V}(\mu_t) + \mathrm{V}(\epsilon_t)$ since ϵ_t and μ_t are uncorrelated. So,

$$\mathrm{V}(\mu_t) = \mathrm{E}\left\{\left[\Psi^{(2)}(B) - 1\right] \upsilon(\mu_t)\right\}.$$

The particular expression for $\upsilon(\mu_t)$ in (2.15) depends on the distribution under investigation from (2.1). For example, in case of Poisson distribution, $\upsilon(\mu_t) = \mu_t$ so that

$$\mathrm{V}(Y_t) = \Psi^{(2)}(B) \, \mathrm{E}(\mu_t) = \Psi^{(2)}(B) \, \beta = \Psi^{(2)}(1) \, \beta,$$

where $\Psi^{(2)}(1) = 1 + \psi_1^2 + \psi_2^2 + \cdots = \sum_{j=1}^{\infty} \psi_j^2$; it can be seen that the variance is constant over t and no additional conditions are required for weak stationarity apart from the usual invertibility of $\Phi(B)$. For other distributions, further invertibility conditions could be required; for example, in the Bernoulli case, even $\Psi^{(2)}(B)$ needs to be invertible to assure stationarity. This proof is due to Benjamin et al. (2003).

We remark that these conditions do not work for other link functions different from the identity; the reason is that, in general, the prediction error in (2.6) $\epsilon_t = h(Y_t) - \eta_t$ is not a MDS (apart from the special case $g \equiv h : identity$).

In order to develop an asymptotic theory for the maximum likelihood estimator (MLE) much more attention has been put in assessing strict stationarity and ergodicity for the GARMA model than proving weak stationarity. For this reason, we will deal with these results in the following section.

2.4.2 M-GARMA

The M-GARMA model (2.7) allows the prediction error to be a MDS. However, the distribution of ϵ_t does depend on $\mathcal{F}_{t-1}$; for this reason, Zheng et al. (2015) pointed out that, in general, the classical condition of invertibility for $\Phi(B)$ is not sufficient for the existence of a stationary distribution of the process $\{g(Y_t)\}_{t\in\mathbb{N}}$. By using the theory of Markov chains, the authors showed that the standard invertibility condition holds only for the special cases in which the link function $\bar{g}(\mu_t) = g(\mu_t) + c$ where c is some function which is constant with respect to μ_t; they call these special cases the *canonical link functions* (a survey of this link function is presented in Zheng et al. (2015)); for the other cases they provided only strict stationarity conditions. However, the authors required $q(y \mid \mathcal{F}_{t-1})$ to be positive everywhere ($\mathbb{R}^+$); this condition is not satisfied for discrete-valued observation process y_t. Thus, their results are valid only for continuous distributions; indeed, in the paper, the attention of the authors is focused on Beta and Gamma distributions.

2.4.3 GLARMA

For the GLARMA models, weak stationarity conditions follow immediately by rewriting (2.8) as a $MA(\infty)$:

$$z_t = \Psi(B)\,\epsilon_t = \sum_{j=1}^{\infty} \psi_j\,\epsilon_{t-j},$$

where the model is initialized by $z_t = 0$ and $\epsilon_t = 0$ for $t \leq 0$. In general, here the process $\{\epsilon_t\}$ is a MDS and in the special case in which Pearson residuals are chosen, it is stationary WN(0,1) and automatically z_t will be (weakly) stationary (and Y_t as well) under usual stationarity and invertibility conditions (roots of $\Phi(B)$ and $\Theta(B)$ lie all outside the unit circle on the complex plan). See Dunsmuir and Scott (2015) for details. Nevertheless, no results are available for strict stationarity apart from the simplest case when $k = 0$, $q = 1$; see Davis et al. (2003), Dunsmuir and Scott (2015), Davis and Liu (2016).

2.5 Strong stationarity

Strong stationarity and ergodicity for models discussed so far are based on several approaches, see Fokianos et al. (2020) for a comprehensive introduction. Here we mainly consider two of them. One is is the perturbation approach introduced by Fokianos et al. (2009) and Fokianos and Tjøstheim (2011), for the linear and log-linear Poisson autoregression models, respectively. The other is the Markov chain theory without irreducibility assumption developed by Matteson et al. (2011), by extending the perturbation argument with Feller properties. These authors showed an application of their approach to the GARMA model as well, see Section 2.5.1. An alternative approach to Markov chain theory without irreducibility assumption is presented by Douc et al. (2013). In this latter paper, an application to the log-linear Poisson autoregression is available, see Section 2.5.2. Similar results are established on the BARMA model, see Moysiadis and Fokianos (2014). For the M-GARMA model only results for continuous variables are available by Zheng et al. (2015). For the GLARMA model, no direct strict-stationarity results have been developed in the literature.

The perturbation approach is an indirect way to establish stability properties of the discrete process $\{Y_t\}$ and it consists of defining a real valued version of the process, by adding a small real perturbation σ to the original process and then showing stochastic properties on the new perturbed process $\left\{Y_t^{(\sigma)}\right\}$. Moreover, it can be proved that, as $\sigma \to 0$, the two processes are arbitrarily close, see the Appendix for details. The Markov chain theory without irreducibility allows to extends results of the perturbation approach to the original process, by exploiting the fact that $\{\mu_t\}$ can be seen as a Markov chain. Showing stationarity and ergodicity for such chain allows one to conclude

for strict stationarity of the integer valued process $\{Y_t\}$. The difference in this approach between Matteson et al. (2011) and Douc et al. (2013) lies only in the additional assumptions required.

We first report an application of the perturbation approach and its extension with Feller properties to the GARMA model in Section 2.5.1. Then, an example of the approach of Douc et al. (2013) to the log-linear Poisson autoregression is presented in Section 2.5.2. We postpone all the theoretical tools required for the application of the two methods in the Appendix.

2.5.1 Strict stationarity and ergodicity for the GARMA model

In this section the conditions under which there exists a strict-sense stationary and ergodic version of the observation process $\{Y_t\}_{t\in\mathbb{N}}$ for the GARMA(1,1) model are given. Define

$$Y_t \mid Y_{0:t-1} \sim q(\mu_t), \tag{2.16}$$

$$g(\mu_t) = \beta + \phi\ \left[g(Y^*_{t-1}) - \beta\right] + \theta\ \left[g(Y^*_{t-1}) - g(\mu_{t-1})\right] \tag{2.17}$$

where Y^*_t is a function which map the value of Y_t to the domain of g. The process $Y_{0:t-1}$ is the set of past values of Y_t from the time 0 until $t-1$. $q(\mu_t)$ is a synthetic notation for (2.1). Three separate cases are considered:

1. $q(\mu)$ is defined for any $\mu \in \mathbb{R}$. In this case the domain of g is $\mathbb{R}$ and $Y^*_t = Y_t$ is taken.
2. $q(\mu)$ is defined for only $\mu \in \mathbb{R}^+$(or μ on any one-sided open interval by analogy). In this case the domain of g is $\mathbb{R}^+$ and $Y^*_t = \max\{Y_t, c\}$ for some $c > 0$ is taken.
3. $q(\mu)$ is defined for only $\mu \in (0, a)$ where $a > 0$ (or any bounded open interval by analogy). In this case the domain of g is $(0, a)$ and $Y^*_t = \min\{\max(Y_t, c), (a - c)\}$ for some $c \in (0, a/2)$ is taken.

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

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

which has the property that $g(\mu) \approx \mu$ for large μ. Examples of valid link functions for Cases 1 and 3 are the identity and logit functions, respectively. Note that model (2.16) is more general than the class of models developed in (2.1) in the sense that it is not necessarily assumed that $q(\cdot)$ belongs to the exponential family.

Perturbed model

The perturbation approach consists of adding a small real-valued perturbation to the discrete-valued time series model in order to obtain a φ-irreducible process (see Definition 1 in the Appendix); then the standard tools for Markov chains could be used to assess stationarity and ergodicity for the perturbed version of the GARMA model. First, ergodicity and stationarity results for the following perturbed model are obtained:

$$Y^{(\sigma)}_t \mid Y^{(\sigma)}_{0:t-1} \sim q(\mu^{(\sigma)}_t)$$

$$g(\mu^{(\sigma)}_t) = \beta + \phi\ \left[g(Y^{(\sigma)*}_{t-1}) - \beta\right] + \theta\ \left[g(Y^{(\sigma)*}_{t-1}) - g(\mu^{(\sigma)}_{t-1})\right] + \sigma Z_{t-1}, \tag{2.18}$$

where $Z_t \sim N(0,1)$ are independent, identically distributed random perturbations, for any $\sigma > 0$, which is a scale factor associated with the perturbation. The value $\mu^{(\sigma)}_0$ is a fixed constant that is taken to be independent of σ, so that $\mu^{(\sigma)}_0 = \mu_0$.

Theorem 1. *The process* $\{\mu_t^{(\sigma)}\}_{t\in\mathbb{N}}$ *specified by the perturbed process* (2.18) *is an ergodic Markov chain and thus is stationary for an appropriate initial distribution for* $\mu_0^{(\sigma)}$*, under the conditions below. This implies that the perturbed process* $\{Y_t^{(\sigma)}\}_{t\in\mathbb{N}}$ *is stationary and ergodic when* $\mu_0^{(\sigma)}$ *is initialized appropriately. The conditions are:*

1. $\mathrm{E}(Y_t^{(\sigma)} \mid \mu_t^{(\sigma)}) = \mu_t^{(\sigma)}$.
2. *(*$2+\delta$ *moment condition): There exist* $\delta > 0$, $r \in [0, 1+\delta)$ *and nonnegative constants* d_1, d_2 *such that*

$$\mathrm{E}(|Y_t^{(\sigma)} - \mu_t^{(\sigma)}|^{2+\delta} \mid \mu_t^{(\sigma)}) \leq d_1 |\mu_t^{(\sigma)}|^r + d_2.$$

3. *g is bijective, increasing, and*
   - 3.1. $g : \mathbb{R} \mapsto \mathbb{R}$ *is concave on* $\mathbb{R}^+$ *and convex on* $\mathbb{R}^-$*, and* $|\phi| < 1$
   - 3.2. $g : \mathbb{R}^+ \mapsto \mathbb{R}$ *is concave on* $\mathbb{R}^+$*, and* $|\phi|, |\theta| < 1$
   - 3.3. $|\theta| < 1$*; no additional conditions on* $g : (0, a) \mapsto \mathbb{R}$.

The proof can be found in the appendix of Matteson et al. (2011). This approach yields stationarity and ergodicity properties for the perturbed model. In order to extend these conclusions to the original unperturbed model the results of the following section are required.

Unperturbed model

In this section, the existence of a stationary distribution for the observation process $\{Y_t\}_{t\in\mathbb{N}}$ of the original (unperturbed) class of GARMA models is proved. Since $\{Y_t\}_{t\in\mathbb{N}}$ is not itself a Markov chain, the existence of a strict-sense stationary ergodic process $\{Y_t\}_{t\in\mathbb{N}}$ is proved by showing that the Markov chain $\{\mu_t\}_{t\in\mathbb{N}}$ has a unique stationary distribution. First, existence of a stationary distribution for the Markov chain is shown by using the weak Feller property. Let $Y_0(x)$ denote the random variable Y_0 conditioned on $\mu_0 = x$. The results of this section are due to Matteson et al. (2011).

Theorem 2. *The process* $\{\mu_t\}_{t\in\mathbb{N}}$ *specified by the GARMA model* (2.17) *has a stationary distribution, and thus is stationary for an appropriate initial distribution for* μ_0*, under the following conditions:*

1. $Y_0(x) \Rightarrow Y_0(x')$ *as* $x \to x'$.
2. $\mathrm{E}(Y_t \mid \mu_t) = \mu_t$.
3. *(*$2+\delta$ *moment condition): There exist* $\delta > 0$, $r \in [0, 1+\delta)$ *and nonnegative constants* d_1, d_2 *such that*

$$\mathrm{E}(|Y_t - \mu_t|^{2+\delta} \mid \mu_t) \leq d_1 |\mu_t|^r + d_2.$$

4. *g is bijective, increasing, and*
   - 4.1. $g : \mathbb{R} \mapsto \mathbb{R}$ *is concave on* $\mathbb{R}^+$ *and convex on* $\mathbb{R}^-$*, and* $|\phi| < 1$
   - 4.2. $g : \mathbb{R}^+ \mapsto \mathbb{R}$ *is concave on* $\mathbb{R}^+$*, and* $|\phi|, |\theta| < 1$
   - 4.3. $|\theta| < 1$*; no additional conditions on* $g : (0, a) \mapsto \mathbb{R}$.

For the proof, Theorem 8 is applied to the chain $\{g(\mu_t)\}_{t\in\mathbb{N}}$ to show that it has a stationary distribution; this implies the same result for the chain $\{\mu_t\}_{t\in\mathbb{N}}$. The state space $S = \mathbb{R}$ of $\{g(\mu_t)\}_{t\in\mathbb{N}}$ is a locally compact complete separable metric space with Borel σ-field. A drift condition for $\{g(\mu_t)\}_{t\in\mathbb{N}}$ is given under the conditions of Theorem 1, for the compact set $A = [-M, M]$ (the drift condition holds when the perturbation $\sigma = 0$). All that remains is to

show that the chain $\{g(\mu_t)\}_{t\in\mathbb{N}}$ is weak Feller. See the Appendix for all the details and definitions. Let $X_t = g(\mu_t)$. For $X_0 = x$ one has that

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

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

Then, uniqueness of the stationary distribution for μ_t is shown, using the asymptotic strong Feller property. It is further assumed that the distribution $\pi_z(\cdot)$ of $g(Y_t)$ conditional on $g(\mu_t) = z$ varies smoothly and not too quickly as a function of z. This mean that $\pi_z(\cdot)$ has the Lipschitz property

$$\sup_{w,z\in\mathbb{R}:w\neq z} \frac{\|\pi_w(\cdot) - \pi_z(\cdot)\|_{TV}}{|w,z|} < B < \infty \tag{2.19}$$

where $\|\cdot\|_{TV}$ is the total variation norm (see Meyn et al. (2009), page 315).

Theorem 3. *Suppose that the conditions of Theorem 2 and the Lipschitz condition* (2.19) *hold, and that there is some $x \in \mathbb{R}$ that is in the support of Y_0 for all values of μ_0. Then there is a unique stationary distribution for $\{\mu_t\}_{t\in\mathbb{N}}$. This implies that $\{Y_t\}_{t\in\mathbb{N}}$ is strictly stationary when μ_0 is initialized appropriately.*

The proof of the theorem can be found in Matteson et al. (2011) and Proposition 8 in Douc et al. (2013).

A similar procedure can be followed to prove strict stationarity and ergodicity for the GARMA model with more than one lag. See Matteson et al. (2011) for further discussion.

2.5.2 Strict stationarity and ergodicity for log-linear Poisson autoregression

The work of Douc et al. (2013) is intended to provide an alternative proof on stationarity and ergodicity for the discrete process Y_t by weaken the Lipschitz assumption (2.19), which is not satisfied for widely used observation-driven models. They specify a wide class of observation-driven model as follows, such as the log-linear Poisson autoregression. Let (X, d) be a locally compact, complete and separable metric space and denote by $\mathcal{X}$ the associated Borel sigma-field. Let $(\mathsf{Y}, \mathcal{Y})$ be a measurable space, H a Markov kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$ and $(x, y) \mapsto f_y(x)$ a measurable function from $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ to $(\mathsf{X}, \mathcal{X})$.

An observation-driven model on $\mathbb{N}$ is a stochastic process $\{(X_t, Y_t), t \in \mathbb{N}\}$ on its space $\mathsf{X} \times \mathsf{Y}$ satisfying the following recursions: for all $t \in \mathbb{N}$,

$$Y_{t+1}|\mathcal{F}_t \sim H(X_t; \cdot), \quad X_{t+1} = f_{Y_{t+1}}(X_t) \tag{2.20}$$

where $\mathcal{F}_t = \sigma(X_l, Y_l; l \leq t, l \in \mathbb{N})$ and $f_{Y_{t+1}}$ is a generic function depending on the observation process $\{Y_l, l \leq t+1\}$. Similarly $\{(X_t, Y_t), t \in \mathbb{Z}\}$ is an observation-driven time series model on $\mathbb{Z}$ if the previous recursion holds for all $t \in \mathbb{Z}$ with $\mathcal{F}_k = \sigma(X_l, Y_l; l \leq t, l \in \mathbb{Z})$.

Denote now by Q the transition probability associated to $\{X_t, t \in \mathbb{N}\}$ defined implicitly by the recursions (2.20). See the Appendix for details. Then, general conditions expressed in terms of H and f are derived under which the processes $\{X_t, t \in \mathbb{N}\}$ and $\{(X_t, Y_t), t \in \mathbb{N}\}$ admit a unique invariant probability distribution.

In the next section we highlight the proof for strict-stationarity and ergodicity for the discrete process. Only the aspects of the proof which significantly different from those in Section 2.5.1 are showed here. We remind the interested reader to the Appendix for the details.

Alternative condition for Markov chain approach without irreducibility

In what follows, if $(\mathsf{E}, \mathcal{E})$ a measurable space, ξ a probability distribution on $(\mathsf{E}, \mathcal{E})$ and R a Markov kernel on $(\mathsf{E}, \mathcal{E})$, denote by P_{ξ}^{R} the probability induced on $(\mathsf{E}^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}})$ by a Markov chain with transition kernel R and initial distribution ξ. Denote by E_{ξ}^{R} the associated expectation. The Lipschitz assumption (2.19) is substituted by

(A3) There exists a kernel $\bar{Q}$ on $(\mathcal{X}^2 \times \{0,1\}, \mathcal{X}^{\otimes 2} \otimes \mathcal{P}(\{0,1\}))$, a kernel $Q^{\sharp}$ on $(\mathcal{X}^2, \mathcal{X}^{\otimes 2})$ and a measurable function $\alpha : \mathsf{X}^2 \to [1, \infty)$ and real numbers $(D, \zeta_1, \zeta_2, \rho) \in (\mathbb{R}^+)^3 \times (0,1)$ such that for all $(x, x') \in \mathsf{X}^2$,

$$1 - \alpha(x, x') \leq d(x, x') W(x, x') \tag{2.21}$$

$$\mathrm{E}_{\delta_x \otimes \delta_{x'}}^{Q^{\sharp}}[d(X_t, X_t')] \leq D \rho^t d(x, x') \tag{2.22}$$

$$\mathrm{E}_{\delta_x \otimes \delta_{x'}}^{Q^{\sharp}}[d(X_t, X_t') W(X_t, X_t')] \leq D \rho^t d^{\zeta_1}(x, x') W^{\zeta_2}(x, x'). \tag{2.23}$$

Moreover, for all $x \in \mathsf{X}$, there exists $\gamma_x > 0$ such that

$$\sup_{x' \in B(x, \gamma_x)} W(x, x') < \infty$$

Some practical conditions for checking (2.22) and (2.23) in (A3) can be denoted.

Lemma 1. *Assume that either (i) or (ii) or (iii) (defined below) holds.*

(i) *There exist $(\rho, \beta) \in (0,1) \times R$ such that for all $(x, x') \in \mathsf{X}^2$*

$$d(X_1, X_1') \leq \rho d(x, x'), \quad \mathrm{P}_{\delta_x \otimes \delta_{x'}}^{Q^{\sharp}} - a.s. \tag{2.24}$$

$$Q^{\sharp} W \leq W + \beta \tag{2.25}$$

(ii) (2.22) *holds and W is bounded.*

(iii) (2.22) *holds and there exist $0 < \alpha < \alpha'$ and $\beta \in \mathbb{R}^+$ such that for all $(x, x') \in \mathsf{X}^2$*

$$d(x, x') \leq W^{\alpha}(x, x')$$

$$Q^{\sharp} W^{1+\alpha'} \leq W^{1+\alpha'} + \beta$$

Then, (2.22) *and* (2.23) *hold.*

All the proof are in the Section 3 of Douc et al. (2013).

The condition (A3) for the Log-linear Poisson autoregression

We now report here the proof of (A3) for the log-linear Poisson autoregression model with one lag. Consider a Markov chain $\{X_t\}_{t \in \mathbb{N}}$ with a transition kernel Q given implicitly by the following recursive equations:

$$\begin{aligned} Y_{t+1} | X_{0:t}, Y_{0:t} &\sim \mathcal{P}(e^{X_t}) \\ X_{t+1} &= d + a\, X_t + b \ln(Y_{t+1} + 1) \end{aligned}$$

where $\mathcal{P}(\lambda)$ is the Poisson distribution with parameter λ. Here $\mathsf{X} = \mathbb{R}$ so $d(x, x') = |x - x'|$ and the function $f_y(x) = d + a\, x + b \ln(1 + y)$. This model called log-linear Poisson autoregression (for details see Fokianos and Tjøstheim (2011)).

Lemma 2. *If $|a + b| \vee |a| \vee |b| < 1$, then (A3) holds.*

Proof. Define $\bar{Q}$ as the transition kernel Markov chain $\{Z_t, t \in \mathbb{N}\}$ with $Z_t = (X_t, X'_t, U_t)$ in the following way. Given $Z_t = (x, x', u)$, if $x \leq x'$, draw independently $Y_{t+1} \sim \mathcal{P}(e^x)$ and $V_{t+1} \sim \mathcal{P}(e^{x'} - e^x)$ and set $Y'_{t+1} = Y_{t+1} + V_{t+1}$. Otherwise, draw independently $Y'_{t+1} \sim \mathcal{P}(e^{x'})$ and $V_{t+1} \sim \mathcal{P}(e^x - e^{x'})$ and set $Y_{t+1} = Y'_{t+1} + V_{t+1}$.

$$
\begin{aligned}
X_{t+1} &= d + a\, x + b \ln(Y_{t+1} + 1), \\
X'_{t+1} &= d + a\, x' + b \ln(Y'_{t+1} + 1), \\
U_{t+1} &= \mathbf{1}_{Y_{t+1} = Y'_{t+1}} = \mathbf{1}_{V_{t+1} = 0}, \\
Z_{t+1} &= (X_{t+1}, X'_{t+1}, U_{t+1})
\end{aligned}
$$

where $\bar{Q}$ satisfies the marginal condition (A-9). Moreover, define for all $x^\sharp = (x, x') \in \mathsf{X}^2$,$Q^\sharp(x^\sharp, \cdot)$ as the law of (X_1, X'_1) where

$$
\begin{aligned}
X_1 &= d + a\, x + b \ln(Y + 1), \quad Y \sim \mathcal{P}(e^{x \wedge x'}), \\
X'_1 &= d + a\, x' + b \ln(Y + 1),
\end{aligned}
\tag{2.26}
$$

and set for all $x^\sharp = (x, x') \in \mathbb{R}^2$,

$$
\alpha(x^\sharp) = \left\{ \exp - e^{x \vee x'} + e^{x \wedge x'} \right\}.
$$

Then, $\bar{Q}$ and $Q^\sharp$ satisfy (A-11). Using twice $1 - e^{-u} \leq u$,it follows that

$$
\begin{aligned}
1 - \alpha(x^\sharp) &= 1 - \left\{ \exp - e^{x \vee x'} + e^{x \wedge x'} \right\} \leq e^{x \vee x'} - e^{x \wedge x'} \\
& \qquad e^{x \vee x'} (1 - e^{-|x - x'|}) \leq W(x, x') |x - x'|
\end{aligned}
$$

with $W(x, x') = e^{|x| \vee |x'|}$ so that (2.21) holds true. To check (2.22) and (2.23), Lemma 1 is applied, by checking option (i). Note first that

$$
\mathrm{P}^{Q^\sharp}_{\delta_x \otimes \delta_{x'}} \{ |X_1 - X'_1| = |a| |x - x'| \} = 1,
\tag{2.27}
$$

so that (2.24) is satisfied. To check (2.25), it can be shown that

$$
\lim_{|x| \vee |x'| \to \infty} \frac{Q^\sharp W(x, x')}{W(x, x')} = 0
\tag{2.28}
$$

and for all $M > 0$,

$$
\sup_{|x| \vee |x'| \leq M} Q^\sharp W(x, x') < \infty
\tag{2.29}
$$

Now, without loss of generality, assume $x \leq x'$. Using (2.26) provides

$$
Q^\sharp W(x, x') = \mathrm{E}\left(e^{|X_1| \vee |X'_1|} \right) \leq \mathrm{E}\left(e^{|X_1|} \right) + \mathrm{E}\left(e^{|X'_1|} \right).
\tag{2.30}
$$

First consider the second term of the right-hand side of (2.30),

$$
\mathrm{E}\left(e^{|X'_1|} \right) \leq e^{|d|} \mathrm{E}(e^{|ax' + b \ln(1+Y)|}).
\tag{2.31}
$$

Noting that if u and v have different signs of if $v = 0$, then $|u + v| \leq |u| \vee |v|$. Otherwise, $|u + v| = (u + v)\mathbf{1}_{v>0} \vee (-u - v)\mathbf{1}_{v<0}$. This implies that

$$
e^{|u+v|} \leq e^{|u|} + e^{|v|} + e^{u+v} \mathbf{1}_{v>0} + e^{-u-v} \mathbf{1}_{v<0}.
$$

and plugging this into (2.31),

$$
\mathrm{E}(e^{|X'_1|}) \leq e^{|d|} \left(e^{|a||x'|} + \mathrm{E}[(1 + Y)^{|b|}] + e^{ax'} \mathrm{E}[(1 + Y)^b] \mathbf{1}_{b>0} + e^{-ax'} \mathrm{E}[(1 + Y)^{-b}] \mathbf{1}_{b<0} \right).
$$

Note that for all $\gamma \in [0,1]$,

$$\mathrm{E}[(1+Y)^\gamma] \leq [\mathrm{E}(1+Y)]^\gamma = (1+e^x)^\gamma \leq 1 + e^{\gamma x} \leq 1 + e^{\gamma x'}.$$

Moreover, since $|b| \in [0,1]$, $b\mathbf{1}_{b>0} \in [0,1]$ and $-b\mathbf{1}_{b<0} \in [0,1]$. Therefore,

$$\begin{aligned}
\mathrm{E}(e^{|X_1'|}) &\leq e^{|d|}\left(e^{|a||x'|} + 1 + e^{|b||x|} + e^{ax'}(1+e^{bx'})\mathbf{1}_{b>0} + e^{-ax'}(1+e^{-bx'})\mathbf{1}_{b<0}\right) \\
&\leq e^{|d|}\left(e^{|a||x'|} + 1 + e^{|b||x|} + e^{|a||x'|} + e^{|a+b||x'|}\right) \\
&\leq e^{|d|}\left(1 + 4e^{\gamma(|x|\vee|x'|)}\right),
\end{aligned}$$

where $\gamma = |a| \vee |b| \vee |a+b| < 1$. The first therm of the right hand side of (2.30) is treated as the second term by setting $x' = x$. So

$$\mathrm{E}(e^{|X_1|}) \leq e^{|d|}\left(1 + 4e^{\gamma(|x|\vee|x'|)}\right),$$

so that using (2.30),

$$Q^\sharp W(x,x') \leq 2e^{|d|}\left(1 + 4e^{\gamma(|x|\vee|x'|)}\right).$$

Since $\gamma \in (0,1)$ and $W(x,x') = e^{|x|\vee|x'|}$, and (2.30) clearly implies (2.28) and (2.29). This proves (A3) and together with (A1)-(A2) provides stationarity conditions for the process $\{Y_t\}$ of the log-linear Poisson autoregression. For further details, see the Appendix. □

For this method the attention is put on showing stability conditions for the model with only one lag. The extension to order greater than the first could be challenging. See Douc et al. (2013).

2.6 Inference

The inferential procedures for observation driven models of discrete processes usually rely on maximum likelihood estimation (MLE). However a misspecified version is available, namely Quasi MLE (QMLE), where the likelihood function considered for the estimation is not necessarily paired with the conditional distribution assumed as a data generating process, see Basawa and Prakasa Rao (1980), Zeger and Liang (1986) and Heyde (1997).

For linear and log-linear Poisson autoregressive time series models, Fokianos et al. (2009) and Fokianos and Tjøstheim (2011) developed maximum likelihood estimation. Quasi-likelihood inference of negative binomial processes has been introduced in Christou and Fokianos (2014). Ahmad and Francq (2016) established consistency and asymptotic normality of the QMLE for the specific case of the Poisson distribution. For the general framework (2.20), Douc et al. (2013) proved the consistency of MLE and QMLE. Asymptotic normality, in the same setting, is later discussed by Douc et al. (2017). Comparable results have been derived by Davis and Liu (2016), based on the approach developed by Neumann (2011). The aim of this section is to give a brief introduction to QMLE for the framework (2.20).

Let (Θ, d) be a compact metric subspace of $\mathbb{R}^p$. Define the parameter vector $\theta \in \Theta$ and the QMLE

$$\hat{\theta}_{n,x} = \arg\max_{\theta\in\Theta} L_{n,x}^\theta \langle Y_{1:n}\rangle, \tag{2.32}$$

with corresponding conditional (quasi) log-likelihood function

$$L_{n,x}^\theta \langle Y_{1:n}\rangle = n^{-1}\log\left(\prod_{t=1}^n h(f^\theta\langle y_{1:t-1}\rangle(x); y_t)\right),$$

where $h(f^\theta\langle y_{1:t-1}\rangle(x); y_t)$ is the density function coming from the kernel H in (2.20) and the notation $f^\theta\langle y_{s:t}\rangle(x) = f_{y_t}^\theta \circ f_{y_{t-1}}^\theta \circ \cdots \circ f_{y_s}^\theta(x)$, $s \leq t$ is the so-called Iterated Random Function (IRF), see Diaconis and Freedman (1999),

with the convention $f^{\theta}\langle y_{1:0}\rangle(x) = x$. Moreover, let $X_0 = x$ be the starting value of the chain X_t in (2.20), then the likelihood is conditional to the starting point x. Here the dependence on the parameter vector θ is emphasized $f^{\theta}_{y_s}(\cdot) = f_{y_s}(\cdot)$.

The following results is due to Douc et al. (2013) and Douc et al. (2017). We make the following assumptions.

(B1) $\{Y_t\}_{t\in\mathbb{Z}}$ is a strict-sense stationary and ergodic stochastic process.

(B2) $\forall (x, y) \in \mathsf{X} \times \mathsf{Y}$, the functions $\theta \mapsto f^{\theta} y(x)$ and $v \mapsto h(v, y)$ are continuous.

(B3) There exists a family of finite random variables $\left\{ f^{\theta}\langle Y_{-\infty:t}\rangle : (\theta, t) \in \Theta \times \mathbb{Z} \right\}$ such that for all $x \in X$,

(i) $\lim_{m\to\infty} \sup_{\theta\in\Theta} d\left[f^{\theta}\langle Y_{-m:0}\rangle(x), f^{\theta}\langle Y_{-\infty:0}\rangle \right] = 0$, a.s.

(ii) $\lim_{t\to\infty} \sup_{\theta\in\Theta} \left| \log h(f^{\theta}\langle Y_{1:t-1}\rangle(x); Y_t) - \log h(f^{\theta}\langle Y_{-\infty:t-1}\rangle; Y_t) \right| = 0$, a.s.

(iii) $\mathrm{E}\left[\sup_{\theta\in\Theta} \left(\log h(f^{\theta}\langle Y_{-\infty:t-1}\rangle; Y_t) \right)_+ \right] < \infty$, where the notation $(\cdot)_+$ is the positive part.

(B4) The true parameter vector $\theta^{\star}$ is assumed to be in Θ^{o}, the interior of Θ.

(B5) The function $\int H(x^{\star}, dy) \log h(x, y)$ has a unique maximum $\{x^{\star}\}$.

Conditions (B1)-(B2) are clearly required so that the estimator $\theta_{n,x}$ is well-defined. Assumption (B3)-(i) assures that, regardless of the initial value of $X_{-m} = x$, the chain X_0 (and thus X_t) can be approximated by a quantity involving the infinite past of the observations. Intuitively, (B3)-(ii) allows the conditional log-likelihood function to be approximated by a stationary sequence involving the infinite past of Y_t. (B3)-(iii) is required in order to obtain a solvable maximization problem and holds for the discrete Y_t (see Remark 18 in Douc et al. (2013)). Assumption (B5) corresponds to an identification condition.

Theorem 4. *Assume that* (B1)-(B5) *hold and* $f^{\theta^{\star}}\langle Y_{-\infty:0}\rangle = f^{\theta}\langle Y_{-\infty:0}\rangle$ *implies that* $\theta = \theta^{\star}$. *Then, for all* $x \in \mathsf{X}$,

$$\lim_{n\to\infty} \hat{\theta}_{n,x} = \theta^{\star}, \quad \textit{a.s.}$$

These results establish strong consistency of the QMLE. For the proof and other details see Douc et al. (2017). An example of derivation of Theorem 4 for the one lag log-linear Poisson AR can be found in Douc et al. (2013). See also Ahmad and Francq (2016), for a similar result.

Finally, the condition under which the QMLE (2.32) is asymptotically normally distributed are investigated. Define the score function

$$\chi^{\theta}(x_t(\theta), y_t) = \nabla_{\theta} x_t(\theta) \frac{\partial \log h(x_t, y_t)}{\partial x_t},$$

and the Hessian matrix

$$K^{\theta}(x_t(\theta), y_t) = \nabla^2_{\theta} x_t(\theta) \frac{\partial \log h(x_t, y_t)}{\partial x_t} + \nabla_{\theta} x_t(\theta) \nabla_{\theta} x_t(\theta)' \frac{\partial^2 \log h(x_t, y_t)}{\partial x_t^2}.$$

Then, define the following notation $f^{\bullet}\langle Y_{-\infty:t-1}\rangle : \theta \mapsto f^{\theta}\langle Y_{-\infty:t-1}\rangle$ and $f^{\bullet}\langle Y_{1:t-1}\rangle(x) : \theta \mapsto f^{\theta}\langle Y_{1:t-1}(x)\rangle$. A further assumption is required.

(B6) : For all $y \in \mathsf{Y}$, the function $v \mapsto h(v, y)$ is twice continuously differentiable. Moreover, there exist $\epsilon > 0$ and a family of a.s. finite random variables

$$\left\{ f^{\theta}\langle Y_{-\infty:t}\rangle : (\theta, t) \in \theta \times \mathbb{Z} \right\}$$

such that $f^{\theta^{\star}}\langle Y_{-\infty:0}\rangle$ is in the interior of X, the function $\theta \mapsto f^{\theta}\langle Y_{-\infty:0}\rangle$ is twice continuously differentiable on some ball $B(\theta^{\star}, \epsilon)$ and for all $x \in \mathsf{X}$,

(i) a.s.,

$$\lim_{t\to\infty} \left\| \chi^{\theta^\star}\left(f^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right) - \chi^{\theta^\star}\left(f^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t\right)\right\| = 0$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^p$.

(ii) a.s.,

$$\lim_{t\to\infty} \sup_{\theta\in B(\theta^\star,\epsilon)} \left\| K^\theta\left(f^\bullet\langle Y_{1:t-1}\rangle(x), Y_t\right) - K^\theta\left(f^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t\right)\right\| = 0$$

where $\|\cdot\|$ denote here any norm on $p\times p$-matrices with real entries.

(iii)

$$\mathrm{E}\left[\left\|\chi^{\theta^\star}\left(f^\bullet\langle Y_{-\infty:0}\rangle, Y_1\right)\right\|^2\right] < \infty, \quad \mathrm{E}\left[\sup_{\theta\in B(\theta^\star,\epsilon)} \left\| K^\theta\left(f^\bullet\langle Y_{-\infty:0}\rangle, Y_1\right)\right\|\right] < \infty$$

Moreover, the matrix

$$\mathcal{J}(\theta_\star) = \mathrm{E}\left[\left(\nabla_\theta g^{\theta_\star}\langle Y_{-\infty:0}\rangle\right)\left(\nabla_\theta f^{\theta_\star}\langle Y_{-\infty:0}\rangle\right)' \frac{\partial^2}{\partial x^2}\log h\left(f^{\theta_\star}\langle Y_{-\infty:0}\rangle, Y_1\right)\right]$$

is non singular.

Intuitively, (B6) assumes that the score function and the information matrix of the data can be approximated by the their counterpart with infinite past of the process. In addition, all of these quantities are assumed to exist.

Theorem 5. *Assume* (B1)-(B6) *hold and* $\hat{\theta}_{n,x} \xrightarrow{p} \theta^\star$. *Then,*

$$\sqrt{n}(\hat{\theta}_{n,x} - \theta^\star) \xrightarrow{d} \mathrm{N}(0, \mathcal{J}(\theta^\star)^{-1}\mathcal{I}(\theta^\star)\mathcal{J}(\theta^\star)^{-1}),$$

where

$$\mathcal{I}(\theta^\star) = \mathrm{E}\left[\left(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle\right)\left(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle\right)'\left(\frac{\partial}{\partial x}\log h\left(f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_1\right)\right)^2\right].$$

The proof relies on the argument of Douc et al. (2017).

Note that, for correctly specified MLE, equation (2.32) is the exact MLE and $\mathcal{J}(\theta^\star) = \mathcal{I}(\theta^\star)$ in Theorem 5, providing the standard ML inference. For further details see Douc et al. (2017). When the quasi-likelihood come from Poisson distribution Ahmad and Francq (2016) proved a similar result for Theorem 5. An analogous conclusion can be found in Christou and Fokianos (2014) for the Negative Binomial distribution.

2.7 Application

The recent outbreak of the new coronavirus called COVID-19 lends itself to a current illustration of the model (2.1, 2.3). The time series we consider is related to the daily number of deaths for COVID-19 in Italy from 21st February 2020 to 20th December 2020. The data can be downloaded by the GitHub repository of the 2019 Novel Coronavirus Visual Dashboard operated by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), `https://github.com/CSSEGISandData/COVID-19`. The time series has a sample size equal to $n = 304$ and is plotted in Figure 2.1, along with its autocorrelation function. The latter shows a temporal correlation spread over several lags in the past. We argue that observation driven models for discrete time series data may be effective in this case. The long time dependence suggests the use of a feedback mechanism, captured by the latent process.

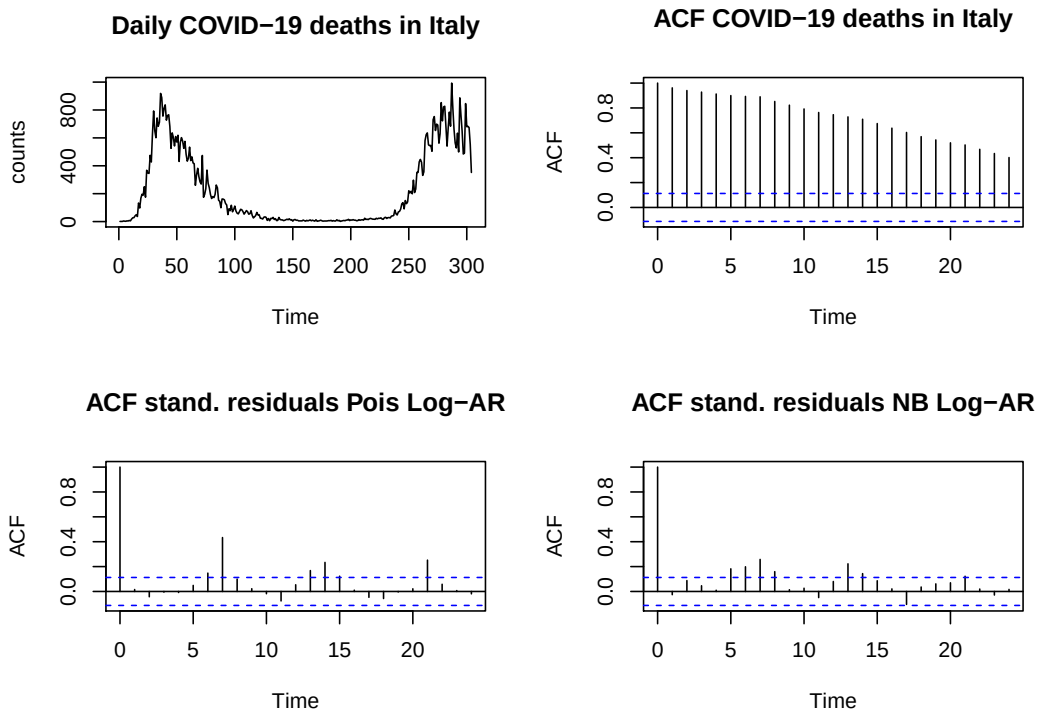

Figure 2.1: Top-left: daily count for COVID-19 deaths in Italy. Top-right: ACF. Bottom-left: ACF standardized residuals for log-AR Poisson model. Bottom-right: ACF standardized residuals for log-AR NB model.

We fit models coming from two different distributions; the Poisson distribution:

$$\mathrm{P}\left(Y_t = y|\mathcal{F}_{t-1}\right) = \frac{\exp(-\mu_t)\mu_t^y}{y!}, \quad y = 0, 1, 2, \ldots$$

and the Negative binomial distribution (NB, henceforth):

$$\mathrm{P}\left(Y_t = y|\mathcal{F}_{t-1}\right) = \frac{\Gamma(\nu + y)}{\Gamma(y+1)\Gamma(\nu)} \left(\frac{\nu}{\nu + \mu_t}\right)^v \left(\frac{\mu_t}{\nu + \mu_t}\right)^y, \quad y = 0, 1, 2, \ldots \tag{2.33}$$

where $\nu > 0$ is the dispersion parameter and μ_t is the conditional expectation; the latter is the same for both distributions. Indeed, equation (2.33) 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, in (2.33), $\mathrm{V}(Y_t|\mathcal{F}_{t-1}) = \mu_t\left(1 + \mu_t/\nu\right) \geq \mu_t$. In the Poisson distribution, the mean and variance are the same.

In order to set a model selection procedure we have estimated the following one-lag models, the log-linear Poisson autoregression (2.11)

$$\log(\mu_t) = \alpha + \phi \log(y_{t-1} + 1) + \gamma \log(\mu_{t-1}),$$

the GARMA model (2.6)

$$\log(\mu_t) = \alpha + \phi \log(y^\star_{t-1}) + \theta\left[\log(y^\star_{t-1}) - \log(\mu_{t-1})\right],$$

where $y^\star_{t-1} = \max\{y_t, c\}$ with $c = 0.1$ and $\alpha = (1-\phi)\beta$ and the GLARMA model (2.7)

$$\log(\mu_t) = \alpha + \gamma \log(\mu_{t-1}) + \theta\left(\frac{y_{t-1} - \mu_{t-1}}{s_{t-1}}\right),$$

where $s_t = \sqrt{\mu_t}$ for the Poisson distribution and $s_t = \sqrt{\mu_t\,(1+\mu_t/\nu)}$ for the NB.

QMLE has been carried out. The log-likelihood function of the Poisson and NB distributions is maximized by using a standard optimizer of `R` based on the BFGS algorithm. The score functions written in terms of predictor $x_t = \log \mu_t$ are:

$$\chi_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\Big(y_t - \exp x_t(\theta)\Big)\frac{\partial x_t(\theta)}{\partial \theta},$$

$$\chi_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{(y_t+\nu)\exp x_t(\theta)}{\exp x_t(\theta)+\nu}\right)\frac{\partial x_t(\theta)}{\partial \theta}.$$

The solution of the system of non-linear equations $\chi_n(\theta) = 0$, if it exists, provides the QMLE of θ (denoted by $\hat{\theta}$). See Section 2.6 for details on the inference. In NB models, the estimation of ν is required. We used the moment estimator, as in Christou and Fokianos (2015):

$$\hat{\nu} = \left\{1/n\sum_{t=1}^{n}\left[(y_t-\hat{\mu}_t)^2 - \hat{\mu}_t\right]/\hat{\mu}_t^2\right\}^{-1},$$

where $\hat{\mu}_t = \mu_t(\hat{\theta})$ from the Poisson model. Clearly, we replace each quantity with the sample counterparts computed at $\hat{\theta}$.

The results of the analysis are summarized in Table 2.1. In the likelihood-based framework, model selection is based on information criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). All the coefficients of the estimation are significant at the usual 5% level. Both AIC and BIC select the NB log-AR model as the best, in the goodness-of-fit sense.

Table 2.1: MLE results for COVID-19 death counts (standard errors in brackets).

Models	$\hat{\alpha}$	$\hat{\phi}$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\nu}$	*AIC*	*BIC*
Pois log-AR	0.154	0.619	0.357	-	-	24.204	35.355
	(0.035)	(0.060)	(0.062)	-			
Pois GARMA	0.211	0.976	-	-0.360	-	24.163	35.314
	(0.036)	(0.006)	-	(0.061)			
Pois GLARMA	0.187	-	0.961	0.038	-	28.047	39.198
	(0.031)	-	(0.008)	(0.003)			
NB log-AR	0.061	0.569	0.424	-	10.733	**15.227**	**26.378**
	(0.023)	(0.036)	(0.035)	-			
NB GARMA	0.157	0.976	-	-0.441	9.123	15.262	26.413
	(0.022)	(0.004)	-	(0.034)			
NB GLARMA	0.712	-	0.822	0.177	4.756	16.636	27.787
	(0.072)	-	(0.016)	(0.011)			

We then assess the adequacy of fit. We check 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 sequence with constant variance. The ACF in our case appears quite uncorrelated for the NB case (see Figure 2.1, for log-AR models).

Another check comes from the probability calibrations, as defined in Gneiting et al. (2007). In particular Czado et al. (2009) introduced a non-randomized version of Probability Integral Transform (PIT) for discrete data. It can

be build by defining the following conditional distribution function

$$F(u|y_t) = \begin{cases} 0, & u \leq P_t(y_t - 1) \\ \frac{u - P_t(y_t - 1)}{P_t(y_t) - P_t(y_t - 1)}, & P_t(y_t) \leq u \leq P_t(y_t - 1) \\ 1, & u \geq P_t(y_t) \end{cases} \tag{2.34}$$

where $P_t(\cdot)$ is the cumulative distribution function at time t (in our case Poisson or NB). If the model is correct, $u \sim Uniform(0, 1)$ and the PIT (2.34) will appear to be the cumulative distribution function of a $Uniform(0, 1)$. The PIT (2.34) is computed for each realisation 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, which are obtained in terms of differences $\bar{F}(\frac{j}{J}) - \bar{F}(\frac{j-1}{J})$ plotted in Figure 2.2. The NB PIT's appear to be closer to $Uniform(0, 1)$, especially for log-linear autoregression and GARMA models.

In order to assess the power of prediction we refer to the concept of sharpness of the predictive distribution defined in Gneiting et al. (2007). It can be measured by some average quantities related to the predictive distribution, which take the form $1/n \sum_{t=1}^{n} d[P_t(y_t)]$, where $d(\cdot)$ is some function called scoring rule. We used some of 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) $-p_t(y_t)/\|p\|$; the ranked probability score (rps) $\sum_{k=0}^{\infty}[P_t(k) - \mathbf{1}(y_t \leq k)]$ and the Dawid–Sebastiani score (dss) $(\frac{y_t - \mu_t}{\sigma_t})^2 + 2 \log \sigma_t$, where μ_t and σ_t are the mean and variance of $P_t(y_t)$. These scores are applied to different models and distributions. The results are summarized in Table 2.2. The NB log-AR model is chosen as the best model, as it has the best predictive performance for all the scoring rules, this confirms the result of the goodness of fit analysis.

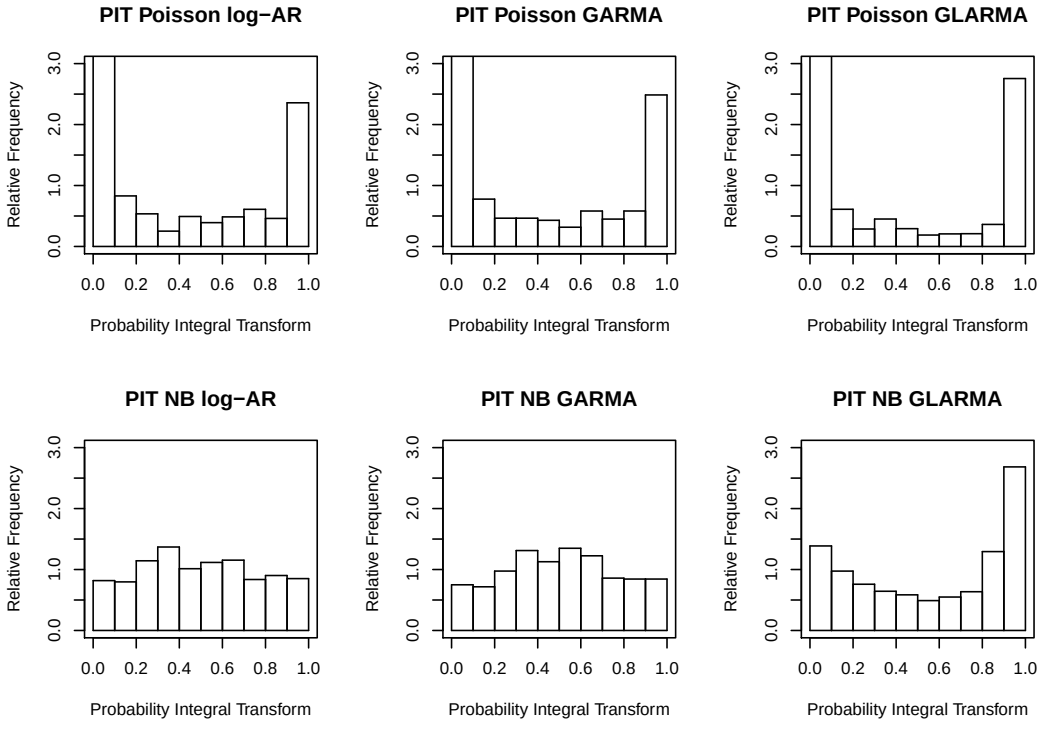

Figure 2.2: Top: PIT's for the Poisson models. Bottom: PIT's for the NB models.

Table 2.2: Predictive performance for COVID-19 death counts (smallest values in bold).

Models	Distribution	logs	qs	sphs	rps	dss
log-AR	Poisson	9.1054	-0.0205	-0.1260	32.6055	21.1890
	NB	**4.6168**	**-0.0324**	**-0.1458**	**29.3324**	**14.0354**
GARMA	Poisson	9.0849	-0.0212	-0.1274	32.5241	21.1019
	NB	4.6345	-0.0320	-0.1448	29.7812	14.1704
GLARMA	Poisson	11.0270	0.0009	-0.0822	36.5751	26.0447
	NB	5.3215	-0.0176	-0.1033	74.0710	16.1614

2.8 Concluding remarks

The most notable observation-driven models for discrete data have been reviewed. The basic stochastic properties required to guarantee their correct use have been presented, as well as the technical tools for their practical application. Increased availability and interest in discrete data encourage the use of these time series models, which will be promising key tools in future works on binary and count data.

For theoretical and substantive reasons, the analysis of discrete-valued times 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 the models that we have discussed 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. Furthermore, novel models not yet specified in the literature could be analyzed in order to obtain better performances in practical applications.

Concerning probabilistic properties, up to the present time, the strict stationarity and ergodicity properties have not been established explicitly for some of the models revised in this chapter (GLARMA and M-GARMA for discrete variables, for example). In principle, the theoretical tools presented in the Appendix would be sufficient to show stability conditions for such models as well as any general framework encompassed in (2.1, 2.3), but the derivations of such stationarity conditions might not be immediate and far from obvious, as shown in Section 2.5 for the GARMA and log-AR models. Then, this would be a useful step further of the literature.

Another aspect which may be interesting to consider is related to the inferential assumptions reported in Section 2.6, which could be generalized to distributions other than Poisson and Negative Binomial and for several different models encompassed in (2.1, 2.3). Lastly, model selection procedures could also be further investigated. We view these aspects as promising topics for future research.

Appendix

Markov chain specification

In order to derive strict stationarity and ergodicity conditions, the problem is rewritten in terms of Markov chain theory. Define an observation-driven model in the most general form:

$$Y_t \mid \mathcal{F}_{t-1} \sim q(\cdot; \mu_t) \tag{A-1}$$

$$\mu_t = c_\delta(Y_{0:t-1}) \tag{A-2}$$

where, henceforth, Y_t indicates the process and y_t its realization. The function q is simply the density function which comes from (2.1) whereas c_δ is some function which describes the form of the dependence from the observation. In

general, $Y_{s:t} = (Y_s, Y_{s+1}, \ldots, Y_t)$ where $s \leq t$. The symbol δ is the vector of parameter of the model. Of course, the initial values $\mu_{0:p-1}$ are supposed to be known. The model in (A-2) can be rewritten as:

$$\mu_t = g_\delta(Y_{t-p:t-1}, \mu_{t-p:t-1}).$$

This way of writing the observation-driven model (Cox (1981)) gives a Markov p-structure for μ_t and then implies that the vector $\mu_{t-p:t-1}$ forms the state of a Markov chain indexed by t. In this case it is possible to prove stationarity and ergodicity of $\{Y_t\}_{t\in\mathbb{N}}$ by first showing these properties for the multivariate Markov chain $\{\mu_{t-p:t-1}\}_{t\geq p}$, then "lifting" the results back to the time series model $\{Y_t\}_{t\in\mathbb{N}}$.

Some useful definition for Markov theorems asserted throughout the paper is introduced here. Define a general Markov chain $X = \{X_t\}_{t\in\mathbb{N}}$ on state space S with σ-algebra $\mathcal{F}$ and define $P^t(x, A) = \mathrm{P}(X_t \in A \mid X_0 = x)$ for $A \in \mathcal{F}$ to be the t-step transition probability starting from state $X_0 = x$.

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

Also, the definition of "aperiodicity" as stated in Meyn et al. (2009) is needed. Define a "period" $d(\alpha) = gcd\{t \geq 1 : P^t(\alpha, \alpha) > 0\}$

Definition 2. *An irreducible Markov chain X is aperiodic if $d(x) \equiv 1$, $x \in X$.*

Definition 3. *A set $A \in \mathcal{F}$ is called a small set if there exists an $m > 1$, a non-trivial measure v on $\mathcal{F}$, and a $\lambda > 0$ such that for all $x \in A$ and all $C \in \mathcal{F}$, $P^m(x, C) \geq \lambda\, v(C)$.*

Now let $\mathrm{E}_x(\cdot)$ denote the expectation under the probability $\mathrm{P}_x(\cdot)$ induced on the path space of the chain defined by $\Omega = \prod_{t=0}^{\infty} X_t$ with respect to $\mathcal{F}^\infty = \bigvee_{t=0}^{\infty} \mathcal{B}(X_t)$ when the initial state $X_0 = x$; where $\mathcal{B}(X_t)$ is the Borel σ-field on X_t.

Theorem 6. *(Drift Conditions). Suppose that $X = \{X_t\}_{t\in\mathbb{N}}$ is φ-irreducible on S. Let $A \subset S$ be small, and suppose that there exist $b \in (0, \infty)$, $\epsilon > 0$, and a function $V : S \to [0, \infty)$ such that for all $x \in S$,*

$$\mathrm{E}_x\left[V(X_1)\right] \leq V(x) - \varepsilon + b\mathbf{1}_{\{x\in A\}}, \tag{A-3}$$

then X is positive Harris recurrent.

The function V is called "Lyapunov function" or "energy function".

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

A stronger form of ergodicity, called "geometric ergodicity", arises if (A-3) is replaced by the condition

$$\mathrm{E}_x\left[V(X_1)\right] \leq \beta V(x) + b\mathbf{1}_{\{x\in A\}} \tag{A-4}$$

for some $\beta \in (0, 1)$ and some $V : S \to [1, \infty)$. Indeed, (A-4) implies (A-3). Eventually, stationarity and ergodicity for the GARMA model would be accomplished if at least one of the sufficient condition (A-3),(A-4) above is fulfilled.

Unfortunately, a problem can occur when the distribution in (A-1) is not continuous (Bernoulli, Poisson,...). In fact, in these cases the Markov chain $\{\mu_{t-p:t-1}\}_{n\geq p}$ may not be φ-irreducible. This occurs whenever Y_t can only take a countable set of values and the state space $\mu_{t-p:t-1}$ is $\mathbb{R}^p$. Then, given a particular initial vector $\mu_{0:p-1}$ the set of possible values for μ_t is countable. Then, Definition 1 is not satisfied. For this reason other theoretical tools are required to solve the problem:

- Perturbation approach
- Feller conditions.

Perturbation approach

First, define the perturbed form of an observation-driven time series model:

$$Y_t^{(\sigma)} \mid Y_{0:t-1}^{(\sigma)} \sim q(\cdot; \mu_t^{(\sigma)}) \tag{A-5}$$

$$\mu_t^{(\sigma)} = g_{\delta,t}(Y_{0:t-1}^{(\sigma)}, \sigma Z_{0:t-1}), \tag{A-6}$$

where $Z_t \sim \phi$ are independent, identically distributed random perturbations having density function ϕ, $\sigma > 0$ is a scale factor associated with the perturbation and $g_{\delta,t}(\cdot, \sigma Z_{0:t-1})$ is a continuous function of $Z_{0:t-1}$ such that $g_{\delta,t}(y, 0) = g_{\delta,t}(y)$ for any y. The value $\mu_0^{(\sigma)}$ is a fixed constant that is taken to be independent of σ, so that $\mu_0^{(\sigma)} = \mu_0$. The perturbed model is constructed to be φ-irreducible, so that one can apply usual drift conditions to prove its stationarity.

Then, it can be proved that the likelihood of the parameter vector δ calculated using (A-6) converges uniformly to the likelihood calculated using the unperturbed model as $\sigma \to 0$. More precisely, the joint density of the observations $Y = Y_{0:t}^{(\sigma)}$ and first t perturbations $Z = Z_{0:t-1}$, conditional on the parameter vector δ, the perturbation scale σ, and the initial value μ_0, is:

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

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

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

Let the likelihood function without the perturbations be denoted by $\mathcal{L}$, so that

$$\mathcal{L}(\delta) = \prod_{k=0}^{t} f\left(Y_k^{(\sigma)}; \mu_k(0)\right).$$

Theorem 7. *Under regularity conditions 1 and 2 below, the likelihood function $\mathcal{L}_\sigma$ based on the perturbed model* (A-5)-(A-6) *converges uniformly on any compact set K to the likelihood function $\mathcal{L}$ based on the original model, i.e.,*

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

for any fixed sequence of observations $y_{0:t}$ and conditional on the initial value μ_0.

So if $\mathcal{L}$ is continuous in δ and has a finite number of local maxima and a unique global maximum on K, the maximum-likelihood estimate of δ based on $\mathcal{L}_\sigma$ converges to that based on $\mathcal{L}$. The proof is in Matteson et al. (2011). Regularity Conditions:

1. For any fixed y the function $q(y; \mu)$ is bounded and Lipschitz continuous in μ, uniformly in $\delta \in K$.
2. For each t, $\mu_t(\sigma Z)$ is Lipschitz in some bounded neighbourhood of zero, uniformly in $\delta \in K$.

Regularity condition 1 holds, e.g., for $q(y; \mu)$ equal to a Poisson or binomial density with mean μ, or a negative binomial density with mean μ and precision parameter φ. $\mu_t(\sigma Z)$ can easily be constructed to satisfy condition 2. One can choose to use the perturbed model (with fixed and sufficiently small perturbation scale σ) instead of the original model, without significantly affecting finite-sample parameter estimates, in order to get the strong theoretical properties associated with stationarity and ergodicity.

Although, it has been shown that the perturbed and original models are closely related, and although one can use drift conditions to show stationarity and ergodicity properties of the perturbed model, this approach does not yield stationarity and ergodicity properties for the original model. In fact, this approach addresses consistency of parameter estimation for the perturbed model when $t \to \infty$ for fixed σ and then shows that as $\sigma \to 0$ the finite sample estimates (for a fixed number of observations t) of the perturbed model approach those of the original one. In order to show real proprieties of the original model one should consider both limits $t \to \infty$ together with $\sigma \to 0$ in which a substantial technical difficulty associated with interchanging the limits arises. For this reason, the Feller properties introduced in the next section are needed.

Feller conditions

To deal with the lack of φ-irreducibility condition, the Feller properties can be used instead.

Definition 4. *A chain evolving on a complete separable metric space S is said to be "weak Feller" if $P(x,\cdot)$ satisfies $P(x,\cdot) \Rightarrow P(y,\cdot)$ as $x \to y$, for any $y \in S$ and where $\Rightarrow$ indicates convergence in distribution.*

In the absence of φ-irreducibility, the "weak Feller" condition can be combined with a drift condition (A-3) or (A-4) to show existence of a stationary distribution (see Tweedie (1988)):

Theorem 8. *Suppose that S is a locally compact complete separable metric space with $\mathcal{F}$ the Borel σ-field on S, and the Markov chain $\{X_t\}_{t\in\mathbb{N}}$ with transition kernel P is weak Feller. Let $A \in \mathcal{F}$ be compact, and suppose that there exist $b \in (0,\infty)$, $\varepsilon > 0$, and a function $V : S \to [0,\infty)$ such that for all $x \in S$, the drift condition* (A-3) *holds. Then there exists a stationary distribution for P.*

Uniqueness of the stationary distribution can be established using the "asymptotic strong Feller" property, defined in Hairer and Mattingly (2006). Before doing it, further definitions are required:

Definition 5. *Let S be a Polish (complete, separable, metrizable) space. A "totally separating system of metrics" $\{d_t\}_{t\in N}$ for S is a set of metrics such that for any $x, y \in S$ with $x \neq y$, the value $d_t(x,y)$ is nondecreasing in t and $\lim_{t\to\infty} d_t(x,y) = 1$.*

Definition 6. *A metric d on S implies the following distance between probability measures μ_1 and μ_2:*

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

where

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

is the minimal Lipschitz constant for ϕ with respect to d.

Definition 7. *A chain is "asymptotically strong Feller" if, for every fixed $x \in S$, there is a totally separating system of metric $\{d_t\}$ for S and a sequence $t_n > 0$ such that*

$$\lim_{\delta\to\infty} \limsup_{t\to\infty} \sup_{y\in B(x,\delta)} \left\| P^{t_n}(x,\cdot) - P^{t_n}(y,\cdot) \right\|_{d_t} = 0$$

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

Definition 8. *A "reachable" point $x \in S$ means that for all open sets A containing x, $\sum_{t=1}^{\infty} P^t(y,A) > 0$ for all $y \in S$.*

Theorem 9. *Suppose that S is a Polish space and the Markov chain $\{X_t\}_{t\in\mathbb{N}}$ with transition kernel P is asymptotically strong Feller. If there is a reachable point $x \in S$ then P can have at most one stationary distribution.*

This is an extension of Hairer and Mattingly (2006). The results of this section lay the foundation for showing convergence and asymptotic properties of maximum likelihood estimators for the discrete-valued observation-driven models.

Coupling construction

Introduce a kernel $\bar{H}$ from $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$ to $(\mathsf{Y}^2, \mathcal{Y}^{\otimes 2})$ satisfying the following conditions on the marginals: for all $(x, x') \in \mathsf{X}^2$ and $A \in \mathcal{Y}$,

$$\bar{H}((x,x');A\times\mathsf{Y}) = H(x,A), \quad \bar{H}((x,x');\mathsf{Y}\times A) = H(x',A). \tag{A-8}$$

Let $C \in \mathcal{Y}^{\otimes 2}$ such that $\bar{H}((x,x');C) \neq 0$ and the chain $\left\{Z_t = (X_t, X_t^{'}, U_t), t \in \mathbb{N}\right\}$ on the"extended" space $(\mathsf{X}^2 \times 0, 1, \mathcal{X}^{\otimes 2} \otimes \mathcal{P}(0,1))$ with transition kernel $\bar{Q}$ implicitly defined as follows. Given $Z_t = (x, x', u) \in \mathsf{X}^2 \times \{0,1\}$, draw $(Y_{t+1}, Y_{t+1}^{'})$ according to $\bar{H}((x,x');\cdot)$ and set

$$X_{t+1} = f_{Y_{t+1}}(x), \quad X_{t+1}^{'} = f_{Y_{t+1}^{'}}(x'),$$

$$U_{t+1} = \mathbf{1}_C(Y_{t+1}, Y_{t+1}^{'}),$$

$$Z_{t+1} = (X_{t+1}, X_{t+1}^{'}, U_{t+1}).$$

The conditions on the marginals of $\bar{H}$, given by (A-8) also imply conditions on the marginals of $\bar{Q}$: for all $A \in \mathcal{X}$ and $z = (x, x', u) \in \mathsf{X}^2 \times \{0,1\}$,

$$\bar{Q}(z; A\times\mathsf{X}\times\{0,1\}) = Q(x,A), \quad \bar{Q}(z;\mathsf{X}\times A\times\{0,1\}) = Q(x',A). \tag{A-9}$$

For $z = (x, x', u) \in \mathsf{X}^2 \times \{0,1\}$, write

$$\alpha(x,x') = \bar{Q}(z;\mathsf{X}^2\times\{1\}) = \bar{H}((x,x');C) \neq 0. \tag{A-10}$$

The quantity $\alpha(x, x')$ is thus the probability of the event $\{U_1 = 1\}$ conditionally on Z_0, taken on $Z_0 = z$. Denote by $Q^\sharp$ the kernel on $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$ defined by: for all $z = (x, x', u) \in \mathsf{X}^2 \times \{0,1\}$ and $A \in \mathcal{X}^{\otimes 2}$,

$$Q^\sharp((x,x');A) = \frac{\bar{Q}(z;A\times\{1\})}{\bar{Q}(z;\mathsf{X}^2\times\{1\})}$$

so that using (A-10),

$$\bar{Q}(z;A\times\{1\}) = \alpha(x,x')\, Q^\sharp((x,x');A). \tag{A-11}$$

This shows that $Q^\sharp((x,x');\cdot)$ is the distribution of $(X_1, X_1^{'})$ conditionally on $(X_0, X_0^{'}, U_1) = (x, x', 1)$.

Assumptions and results of the alternative Markov chain approach

Consider the following assumptions.

(A1) The Markov kernel Q is weak Feller. Moreover, there exist a compact set $C \in \mathcal{X}$,$(b, \varepsilon) \in \mathbb{R}_*^+ \times \mathbb{R}_*^+$ and a function $V : \mathsf{X} \to \mathbb{R}^+$ such that

$$QV \leq V - \varepsilon + b\mathbf{1}_C.$$

(A2) The Markov kernel Q has reachable point.

Assumption (A1) implies, by Tweedie (1988), that the Markov kernel Q admits at least one stationary distribution. Assumptions (A2)-(A3) are then used to show that this stationary distribution is unique.

Note that assumptions (A1)-(A2) are the same of Theorem 8 and 9 and they can be proved for each observation driven model as has been done for the GARMA model; assumption (A3) weakens the Lipschitz condition (2.19) by introducing a function W in (2.21). This allows to treat models which do not satisfy the Lipschitz condition (2.19); for example the log-linear Poisson autoregression (see Section below).

Theorem 10. *Assume that (A1)-(A3) hold. Then, the Markov kernel Q admits a unique invariant probability measure.*

Proposition 1. *Assume that the Markov kernel Q admits a unique invariant probability measure. Then, there exists a strict-sense stationary ergodic process on $\mathbb{Z}$, $\{Y_t\}_{t\in\mathbb{Z}}$, the solution to the recursion* (2.20).

These results can be found in Douc et al. (2013).

Computational aspects

The replication code for the application in Section 2.7 is available at `https://github.com/mirkoarmillotta/covid_code`. First, a function for the log-likelihood and the gradient of the log-linear Poisson autoregression is provided. The code for the other models works in a similar way and it is available upon request. Then, a function to perform the QMLE is presented. Finally, we give the code for the COVID-19 example and the relative plots. The code to perform the PIT is due to Czado et al. (2009) and it is available in the reference therein.

Bibliography

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

Al-Osh, M. and A. A. Alzaid (1987). First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis 8*, 261–275.

Alzaid, A. and M. Al-Osh (1990). An integer-valued pth-order autoregressive structure (INAR (p)) process. *Journal of Applied Probability*, 314–324.

Basawa, I. V. and B. L. S. Prakasa Rao (1980). *Statistical Inference for Stochastic Processes*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London-New York. Probability and Mathematical Statistics.

Benjamin, M., R. Rigby, and D. Stasinopoulos (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association 98*(461), 214–223.

Billingsley, P. (1995). *Probability and Measure* (3 ed.). Wiley.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*(3), 307–327.

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

Box, G. E. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Prentice-Hall Inc.

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

Christou, V. and K. Fokianos (2015). On count time series prediction. *Journal of Statistical Computation and Simulation 85*(2), 357–373.

Clark, N. J., M. S. Kaiser, and P. M. Dixon (2018). A spatially correlated auto-regressive model for count data. *arXiv preprint arXiv:1805.08323*.

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

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*(5), 777–795.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics 65*(4), 1254–1261.

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

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

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

Diaconis, P. and D. Freedman (1999). Iterated random functions. *SIAM review 41*(1), 45–76.

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

Douc, R., K. Fokianos, and E. Moulines (2017). Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electronic Journal of Statistics 11*, 2707–2740.

Dunsmuir, W. and D. Scott (2015). The GLARMA package for observation-driven time series regression of counts. *Journal of Statistical Software 67*(7), 1–36.

Engle, R. F. (1982, 06). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*(4), 987–1007.

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

Fokianos, K., B. Kedem, et al. (2003). Regression theory for categorical time series. *Statistical Science 18*(3), 357–376.

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

Fokianos, K., B. Støve, D. Tjøstheim, and P. Doukhan (2020). Multivariate count autoregression. *Bernoulli 26*, 471–499.

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

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 243–268.

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 (Statistical Methodology) in press*.

Hairer, M. and J. C. Mattingly (2006). Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. *Annals of Mathematics 164*(3), 993–1032.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Cambridge University Press.

Heyde, C. C. (1997). *Quasi-likelihood and its Application*. Springer Series in Statistics. Springer-Verlag, New York. A General Approach to Optimal Parameter Estimation.

Ho, S.-L., M. Xie, and T. N. Goh (2002). A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. *Computers & Industrial Engineering 42*(2-4), 371–375.

Kauppi, H. and P. Saikkonen (2008). Predicting U.S. recessions with dynamic binary response models. *The Review of Economics and Statistics 90*(4), 777–791.

Li, W. K. (1994). Time series models based on generalized linear models: some further results. *Biometrics 50*(2), 506–511.

Matteson, D. S., D. B. Woodard, and S. G. Henderson (2011). Stationarity of generalized autoregressive moving average models. *Electronic Journal of Statistics 5*, 800–828.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition* (2 ed.). Chapman & Hall.

Meyn, S., R. L. Tweedie, and P. W. Glynn (2009). *Markov Chains and Stochastic Stability* (2 ed.). Cambridge University Press.

Moysiadis, T. and K. Fokianos (2014). On binary and categorical time series models with feedback. *Journal of Multivariate Analysis 131*, 209–228.

Neumann, M. H. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli 17*(4), 1268–1284.

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

Sen, P., M. Roy, and P. Pal (2016). Application of ARIMA for forecasting energy consumption and GHG emission: a case study of an Indian pig iron manufacturing organization. *Energy 116*, 1031–1038.

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

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association 89*(425), 208–218.

Tong, H. and K. Lim (1980). Threshold autoregression, limit cycles and cyclical data-with discussion. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 42*(3), 245–292.

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

Walker, G. T. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 131*(818), 518–532.

Wang, Y., J. Wang, G. Zhao, and Y. Dong (2012). Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: a case study of China. *Energy Policy 48*, 284–294.

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, Containing Papers of a Mathematical or Physical Character 226*(636-646), 267–298.

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

Zeger, S. L. and K.-Y. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121–130.

Zheng, T., H. Xiao, and R. Chen (2015). Generalized ARMA models with martingale difference errors. *Journal of Econometrics 189*(2), 492 – 506.

Chapter 3

Observation driven models for discrete-valued time series

Mirko Armillotta[1], Alessandra Luati[1] and Monia Lupparelli[2]

[1] *Department of Statistical Sciences, University of Bologna, 41 st. Belle Arti, 40126, Bologna, Italy.*
Email: mirko.armillotta2@unibo.it, alessandra.luati@unibo.it

[2] *Department of Statistics, Computer Science, Applications, University of Florence, 59 ave. Morgagni, 50134, Florence, Italy.*
Email: monia.lupparelli@unifi.it

Abstract

Statistical inference for discrete-valued time series has not been developed as 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, it is quite arduous to derive stochastic properties which simultaneously hold across different specifications. In this paper, inference for a general class of 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 unified framework 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. Two empirical applications are also discussed, for count data. The first application is a novel application to hurricane data in the North Atlantic Basin; the second concerns time series on the spread of an infection.

Keywords: count data, generalized ARMA models, likelihood inference, link function.

3.1 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 Yule (1927) and Walker (1931), who introduced the concept of autoregression, a dynamic model for the conditional mean of a stochastic process. In the same years, Slutsky (1927, 1937) 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 Box and Jenkins (1970, 1976) 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 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 since 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 (Startz, 2008) and the Wold representation has no direct interpretation, see the discussion in the recent handbook edited by Davis et al. (2016). Since the AR- and MA-like models first introduced by Zeger and Qaqish (1988) and Li (1994), there have been some relevant specifications, such as the generalized ARMA (GARMA) by Benjamin et al. (2003) and their martingalised version, the M-GARMA by Zheng et al. (2015), as well as the generalized linear ARMA (GLARMA) by Davis et al. (2003). An interesting class of autoregression models for count data has been proposed by Fokianos et al. (2009) and Fokianos and Tjøstheim (2011), inspired to the generalized linear transformation of McCullagh and Nelder (1989). Integer-valued time series with extreme observations have been recently dealt with by Gorgi (2020), based on the beta-negative binomial distribution.

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 and even to include further new specifications. 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 is also in contrast with the growing attention, in recent years, to high dimensional data sets involving dynamic binary and count data, in different contexts, such as the number of clicks or amount of intra-day stock transactions (Davis and Liu, 2016; Ahmad and Francq, 2016). Attempts in this direction have been made by Douc et al. (2013) 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 Douc et al. (2013) 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.

If we were like to summarise the main results developed in the literature, on the side of the stochastic properties, Matteson et al. (2011) 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 Neumann (2011), based on contraction conditions, and in Doukhan et al. (2012), based on the weak dependence approach. Fokianos et al. (2009) and Fokianos and Tjøstheim (2011) 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 Christou and Fokianos (2014) under the assumption of a Negative Binomial distribution as the data generating process.

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. Douc et al. (2013) prove the consistency of MLE and QMLE for the general framework they proposed. Asymptotic normality, in the same setting, is later discussed by Douc et al. (2017). Comparable results have been derived by Davis and Liu (2016), based on the approach developed by Neumann (2011), and by Ahmad and Francq (2016) 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 observation driven model for discrete-valued stochastic processes that encompasses the existing models in literature and includes novel specifications. In the terminology of Cox (1981), observa-

tion 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 class of 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 methodological contribution of the paper consists in the development of the stochastic theory and the likelihood inference holding for all the models in the class, through a non-trivial extension of the theory of Matteson et al. (2011) as far as stationarity and ergodicity are concerned, and of the theory of Douc et al. (2013) and Douc et al. (2017) 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, we introduce a general modelling framework 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 which hold simultaneously for the entire class of models (strict stationarity and ergodicity), (iv) to implement quasi-maximum likelihood (QMLE) inference which also allows us to define model selection criteria across different, and not nested, models, (v) to derive the asymptotic properties of QMLE, (vi) to make all the models encompassed in the framework fully applicable in practice.

On the side of applications, the analysis of two real datasets is performed, for count time series. The first is a novel application to hurricane data in the North Atlantic Basin. It is well-established that warming earth should experience more hurricanes and/or stronger individual storms. For this reason, forecasting annual hurricane counts is of great interest and several Poisson-based models have been developed; see Xiao et al. (2015) and references therein. More recently, Livsey et al. (2018) used autoregressive fractionally integrated moving average models to construct a Poisson model able to capture the long-range effect for the hurricane trend. Given the short length of the data record (49 years), their model based on a generalization of fractionally integration methodology to discrete data cannot properly address this issue. Nevertheless, the Poisson dynamics seems to be not always suitable and further models for over-dispersed count distributions have been proposed founded on negative binomial assumptions (Villarini et al., 2010). Models included in the general framework are used for the analysis of hurricane data in the North Atlantic Basin considering both the Poisson and negative binomial assumption for the generating process. We pay specific attention to model selection which is performed by using information criteria that also accounts for model misspecification. With the focus on model comparison, the second application uses a test-bed time series in count data analysis, on the spread of an infection, Escherichia coli, in the German region of North-Rhine Westphalia.

3.2 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, $s \leq t$. The process Y_t is adapted to the filtration $\mathcal{F}$ and $\mathrm{E}|Y_t| < \infty$ 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 μ_t, is a member of the one-parameter exponential family

$$q(Y_t|\mathcal{F}_{t-1}) = \exp\left\{Y_t\, f(X_t) - A\left(X_t\right) + d(Y_t)\right\}, \tag{3.1}$$

$$X_t = g(\mu_t) = \mathbf{Z}_t^T\boldsymbol{\alpha} + \sum_{j=1}^{k}\gamma_j g(\mu_{t-j}) + \sum_{j=1}^{p}\phi_j h(Y_{t-j}) + \sum_{j=1}^{q}\theta_j\left[\frac{h(Y_{t-j}) - \bar{g}(\mu_{t-j})}{\nu_{t-j}}\right], \tag{3.2}$$

where it is assumed that 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 link function. The function $A(\cdot)$ (log-partition) and $d(\cdot)$ are specific functions which define the particular distribution (McCullagh and Nelder, 1989). The mapping $f(\cdot)$ is a twice-differentiable bijective function, chosen according to the model of interest. Each exponential family in the form (3.1) can be re-parametrised in the canonical form:

$$q(Y_t|\mathcal{F}_{t-1}) = \exp\left\{Y_t\, Q_t - \bar{A}\,(Q_t) + d(Y_t)\right\}\,, \tag{3.3}$$

where the sequence $Q_t = f(X_t) = f[g(\mu_t)] = \tilde{f}(\mu_t)$ is called 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-parametrisation of $A\,(\cdot)$ with respect to Q_t. It is known that for the exponential family (3.3) 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 (3.1), with $f(X_t) = X_t$, so that (3.1) and (3.3) 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 make some examples to clarify the nature of the framework.

Example 3. *In the setting (3.1, 3.2), 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* (3.2) *becomes a log-linear model on the response* $\log(\mu_t)$. *It is possible to model* (3.2) *with a different shape of* $g(\cdot)$*; for example, one may be interested to 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* (3.1), *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 4. *The Gaussian distribution (with known variance) is obtained by setting* $f(X_t) = X_t$, $g(\mu_t) = \frac{\mu_t}{\sigma_t^2}$, $A\,[g(\mu_t)] = \frac{\mu_t^2}{2\sigma_t^2}$ *and* $d(Y_t) = \log\left[-\frac{1}{\sqrt{2\pi\sigma_t^2}}\exp\left(-\frac{Y_t^2}{2\sigma_t^2}\right)\right]$. *One can verify that* $\mu_t = \sigma_t^2 X_t$, *so* $A(X_t) = \sigma_t^2 X_t^2/2$, *with first and second derivatives* μ_t *and* σ_t^2, *respectively.*

Note that the process $\{Y_t\}_{t\in T}$ is observed whereas $\{\mu_t\}_{t\in T}$ is not. However, from equation (3.2), 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 also the reason why we refer to "observation-driven models". The function $h(Y_t)$ is called "data-link function" since it is applied to the process Y_t whereas $\bar{g}(\mu_t)$ is said "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 (3.2) and the distribution of interest in equation (3.1). We define the prediction error as the ratio

$$\varepsilon_t = \frac{h(Y_t) - \bar{g}(\mu_t)}{\nu_t} \tag{3.4}$$

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, (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, in symbols $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]$, 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 (3.4) form a white noise (WN) sequence, with unit variance.

The vector $\mathbf{Z}_t = [1, Z_{1t}, \dots, Z_{st}]^T$ in equation (3.2) 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 $Y_{t-j-1}, \dots$); finally, θ_j represents the analogous of a moving average component, since the ratio (3.4) can be built so as to have an error-type behaviour. In general, all the functions involved are not constrained to assume the same shape and the additive parts of the model (3.2) can be arranged in different ways. Clearly, sub-models are allowed. This leads to a quite general and flexible framework which encompasses the most frequently used models for discrete-valued observation processes and also new ones.

3.2.1 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, (Benjamin et al., 2003). Here, the distribution of the process is usually assumed to be the one-parameter exponential family (3.1). From equation (3.2) the GARMA model is obtained when $k = 0$, by setting $g \equiv \bar{g} \equiv h$ and $\nu_t = 1$, so that,

$$g(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{p} \phi_j g(Y_{t-j}) + \sum_{j=1}^{q} \theta_j \left[g(Y_{t-j}) - g(\mu_{t-j}) \right] , \tag{3.5}$$

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 Benjamin et al. (2003).

A suitable extension of the GARMA model (3.5), the martingalised GARMA (M-GARMA), has recently been introduced by Zheng et al. (2015); it is derived from (3.2) by setting $k = 0$, $g(\mu_t) \equiv \bar{g}(\mu_t) = \mathrm{E}[h(Y_t)\,|\mathcal{F}_{t-1}]$ and $\nu_t = 1$:

$$\bar{g}(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{p} \phi_j h(Y_{t-j}) + \sum_{j=1}^{q} \theta_j \left[h(Y_{t-j}) - \bar{g}(\mu_{t-j}) \right] . \tag{3.6}$$

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 Shephard (1995), Rydberg and Shephard (2003) and Davis et al. (2003) with the name of Generalized Linear Autoregressive Moving Average model (GLARMA); here again the distribution is the exponential family (3.1). We can write the GLARMA model (3.2) by setting $p = 0$, h as the identity and $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)\,|\mathcal{F}_{t-1}] = \mathrm{E}(Y_t\,|\mathcal{F}_{t-1}) = \mu_t$:

$$g(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j g(\mu_{t-j}) + \sum_{j=1}^{\tilde{q}} \theta_j \varepsilon_{t-j} , \tag{3.7}$$

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 $j = 1, \dots, \tilde{q}$, where τ_j is a free parameter. The formulation of the constant term in equation (3.7) as a function of $\boldsymbol{\beta}$ is equivalent to equation (13) in Dunsmuir and Scott (2015), the alternative definition of the GLARMA model originally introduced in Davis et al. (2003). Note that here, if $\nu_t = \sigma_t$, then the prediction error $\varepsilon_t = \frac{Y_t - \mu_t}{\nu_t}$ is a white noise process with unit variance.

Another promising stream of literature is due to Fokianos et al. (2009), who introduced Poisson autoregression, henceforth Pois AR, which is obtained when (3.1) is $Pois(\mu_t)$, with $f(X_t) = \log(X_t)$, and in equation (3.2), we have $q = 0$ and $g \equiv h : identity$:

$$\mu_t = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \mu_{t-j} + \sum_{j=1}^{p} \phi_j Y_{t-j} . \tag{3.8}$$

The parameters in equation (3.8) are constrained in the positive real line. A variant of (3.8) is the log-linear Poisson autoregression, henceforth Pois log-AR, (Fokianos and Tjøstheim, 2011) which is obtained by (3.2) when $q = 0$, $f(X_t) = X_t$, $g(\mu_t) = \log(\mu_t)$ and $h(Y_t) = \log(Y_t + 1)$

$$\log(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \log(\mu_{t-j}) + \sum_{j=1}^{p} \phi_j \log(Y_{t-j} + 1)\,. \tag{3.9}$$

For Poisson data, the GARMA model (3.5) 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 (3.8) or (3.9). A model like (3.9) could be used also for Negative Binomial data, by rewriting the distribution in terms of the expected value parameter μ_t (Christou and Fokianos, 2014):

$$q\left(Y_t|\mathcal{F}_{t-1}\right) = \frac{\Gamma(\nu + Y_t)}{\Gamma(Y_t + 1)\Gamma(\nu)} \left(\frac{\nu}{\nu + \mu_t}\right)^{v} \left(\frac{\mu_t}{\nu + \mu_t}\right)^{Y_t} \tag{3.10}$$

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 (3.10) with model (3.9) is obtained from the distribution (3.1), 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 BARMA model (Li (1994); Startz (2008)), introduced for Binomial data, is obtained when (3.1) is $Bin(a, \mu_t)$, where a is known and the probability parameter $p_t = \mu_t/a$, and, in (3.2), $\gamma = 0$, $h : identity$ ($\bar{g}(\mu_t)$ reduces to μ_t) and $c = 0$. Then

$$g(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j \left[Y_{t-j} - \mu_{t-j}\right]\,. \tag{3.11}$$

Even if, this model is thought for Binomial distribution, so typically $g : logit$ or $g : probit$, in general, the link function g can be any suitable function.

3.2.2 New model specifications

Other models of potential interest not explicitly included in the existent literature are indeed encompassed in the framework (3.1)-(3.2). We discuss a class of $g$link-ARMA models. As relevant instance consider the log-ARMA model

$$\log(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \log(\mu_{t-j}) + \sum_{j=1}^{p} \phi_j \log(Y_{t-j} + 1) + \sum_{j=1}^{q} \theta_j \left[\frac{\log(Y_{t-j} + 1) - \bar{g}(\mu_{t-j})}{\nu_{t-j}}\right] \tag{3.12}$$

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 (3.12) detects the autoregressive effect of the past lags of Y_t, but it also accounts for a long past feedback effect, via lags of μ_t; then, a white noise prediction error $\varepsilon_t = \left[\frac{\log(Y_t+1)-\bar{g}(\mu_t)}{\nu_t}\right]$ is added to the functional transformation of the data, where $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{V}(\varepsilon_t) = 1$. The same model (3.12), when (3.1) is $Bin(a, \mu_t)$, is resorted 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 $d(Y_t) = \log\binom{a}{Y_t}$. On the same line, a logit-ARMA model can be specified for Binomial data as a combination of the BARMA model from Li (1994) and an autoregressive component:

$$\log\left(\frac{\mu_t}{a - \mu_t}\right) = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \log\left(\frac{\mu_{t-j}}{a - \mu_{t-j}}\right) + \sum_{j=1}^{p} \phi_j\, Y_{t-j} + \sum_{j=1}^{q} \theta_j \left[\log(Y_{t-j} + 1) - \bar{g}(\mu_{t-j})\right] \tag{3.13}$$

where, in equation (3.1) 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 be specified also by replacing the $logit$ function with the $probit$ link function.

The usefulness of the specifications (3.12)-(3.13) 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, when the distribution of $Y_t|\mathcal{F}_{t-1}$ is Log-normal(μ_t, σ^2), the expectation $\bar{g}(\mu_t) = \mathrm{E}\left[\log(Y_t+1)|\mathcal{F}_{t-1}\right] = \log(\mu_t) - 1/2\sigma^2$. For a comprehensive discussion on the closed form solutions see Zheng et al. (2015). In the case of Binomial or Poisson data, though, such closed forms are not available and it seems reasonable to use an approximation from the Taylor expansion around the mean μ_t, like $\bar{g}(\mu_t) = \mathrm{E}\left[h(Y_t)|\mathcal{F}_{t-1}\right] \approx h(\mu_t)$. However, this would reduce models (3.12)-(3.13) to a reparametrized version of the already showed log-AR model described in equation (3.9). Despite the wide use of the Poisson model for count data and the default negative Binomial alternative to account for overdispersion, both choices fail when data present underdispersion or an excess of zero value observations (Englehardt et al., 2012). For instance, the use of the discrete Weibull distribution of Nakagawa and Osaki (1975) and its generalizations are quite popular in these contexts; see Peluso et al. (2019) for a discussion. The generalization of distributions to accommodate specific data structure represents an active research area which may benefit from a flexible specification of $g$link-ARMA type models.

Furthermore, novel and potentially useful models also arise when equation (3.2) involves the use of a Box-Cox transformation (Box and Cox, 1964):

$$\frac{\mu_t^\lambda - 1}{\lambda} = \mathbf{Z}_t^T \boldsymbol{\alpha} + \sum_{j=1}^{k} \gamma_j \frac{\mu_{t-j}^\lambda - 1}{\lambda} + \sum_{j=1}^{p} \phi_j \frac{Y_{t-j}^\lambda - 1}{\lambda} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}\,, \tag{3.14}$$

where $g(z) = h(z) = \frac{z^\lambda - 1}{\lambda}$, $\varepsilon_t = \frac{\lambda\left[Y_t^\lambda - \mathrm{E}(Y_t^\lambda|\mathcal{F}_{t-1})\right]}{\mathrm{V}(Y_t^\lambda|\mathcal{F}_{t-1})}$, by equation (3.4) and λ is the transformation parameter, which can be chosen according to some estimation procedure, such as profile likelihood. Note that when $\lambda = 0$ the model (3.14) reduces to model (3.12) with $\log(Y_{t-j})$ instead of $\log(Y_{t-j}+1)$. This model can exploit the usefulness of the Box-Cox transformation, possibly leading to a more stable variance and improving symmetry of the distribution. However, the link function $g(\mu_t) = \frac{\mu_t^\lambda - 1}{\lambda}$ is not canonical for any distribution encompassed in the exponential family (3.1), hence the function $f(\cdot)$ needs to be chosen according to the conditional distribution of Y_t.

3.3 Stochastic properties

This section provides the conditions for the discrete-valued stochastic process $\{Y_t\}_{t\in T}$ to be stationary and ergodic by using Markov chain theory. Although $\{Y_t\}_{t\in T}$ is not itself a Markov chain, the process $\{\mu_t\}_{t\in T}$ is. Then, by proving that the chain $\{\mu_t\}_{t\in T}$ has a unique invariant distribution, one also has that the double sequence $\{Y_t, \mu_t\}_{t\in T}$ is a Markov chain with unique distribution. Hence, the process $\{Y_t\}_{t\in T}$ is stationary and ergodic, see Matteson et al. (2011) and Douc et al. (2013).

3.3.1 Stationarity and ergodicity

The proof of the stability conditions is established by showing the ergodicity of a first order Markov chain process (see below). Since this approach is usually challenging beyond the order one chain, we set (3.2) with $k = p = q = 1$, in the absence of covariates ($\mathbf{Z}_t^T\boldsymbol{\alpha} = \alpha$) and with unitary scaling sequence, $\nu_t = 1$ for $t \in T$:

$$g(\mu_t) = \alpha + \gamma\, g(\mu_{t-1}) + \phi\, h(Y_{t-1}^*) + \theta\left[h(Y_{t-1}^*) - \bar{g}(\mu_{t-1})\right]\,, \tag{3.15}$$

where the function Y_t^* modifies the values of Y_t to lie into the domain of $h(\cdot)$. In Remark 2 we discuss an extension which includes the scaling sequence. In the first order observation-driven model (3.15) the series μ_t can be determined recursively by knowing the starting point μ_0 and the observations $Y_0, \ldots, Y_{t-1}$. 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:

1. $q(Y_t|\mathcal{F}_{t-1})$ for $\mu \in \mathbb{R}$. The domain of g and h is $\mathbb{R}$ and $Y_t^* = Y_t$.
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 $c \geq 0$.
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 $P^t(x, A) = \mathrm{P}(X_t \in A \mid X_0 = x)$ 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 $l_1, l_2 \geq 0$ such that $\mathrm{E}(|Y_t - \mu_t|^{2+\delta} \mid \mu_t) \leq l_1 |\mu_t|^r + l_2$.
(A3) g and h are bijective, increasing and

1. If $\bar{g}(\mu_t) = g(\mu_t)$,
   1.1. $h : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $g : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $|\gamma| + |\phi| < 1$
   1.2. $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $(|\gamma| + |\phi|) \vee |\gamma - \theta| < 1$
   1.3. $h : (0, a) \mapsto \mathbb{R}$ 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 $L \leq 1$,
   2.1. $h : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $g : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $|\gamma| + |\phi| < 1$
   2.2. $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $|\gamma| + (|\phi| \vee |\theta|) < 1$
   2.3. $h : (0, a) \mapsto \mathbb{R}$ 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 11. *Suppose that* (A1)-(A4) *hold. Then, the process* $\{\mu_t\}_{t\in T}$ *in* (3.15) *has a unique stationary distribution. This implies that* $\{Y_t\}_{t\in T}$ *is strict-sense stationary and ergodic.*

The proof is postponed in the Supplementary Materials and is carried out by showing that the Markov chain $\{X_t\}_{t\in T}$ has a unique stationary distribution, under the conditions of Theorem 11. This is done by proving a drift condition for the chain which is sufficient for φ-irreducible Markov chains (Meyn et al., 2009). However the discreteness of $\{Y_t\}_{t\in T}$ may lead to a non-φ-irreducible chain. Indeed, the process X_t depends on values of Y_t, hence, it lies in a countable subset of S, which implies the non φ-irreducibility of the chain. Therefore, by following the Markov chain theory without irreducibility assumption (Matteson et al., 2011; Douc et al., 2013), the weak Feller and the asymptotic strong Feller properties are required on the chain X_t, providing the desired result.

Assumption (A1) automatically holds when $\mu_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$, as in the case of equation (3.1). For model (3.15), the σ-algebra generated by μ_t is a subset of $\mathcal{F}_{t-1}$, and for the tower property $\mathrm{E}(Y_t|\mu_t) = \mathrm{E}[\mathrm{E}(Y_t|\mathcal{F}_{t-1})|\mu_t] = \mu_t$. Assumption (A2) is a mild moment condition generally satisfied for usual discrete distributions (Poisson, Binomial); see Matteson et al. (Cor.6,7, 2011) for details.

Remark 1. *It is worth noting that Theorem 11 is not restricted to distribution* (3.1) *since it involves only the moment conditions in assumptions* (A1)-(A2)*.*

The conditions on the shape of the link functions g and h in (A3) are quite standard. While Assumption (A4) might be not immediate to verify, it can usually be replaced with an alternative condition, which is easier to check:

(A5) The distribution (3.1) 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 by Matteson et al. (2011) for the Poisson and Binomial distribution; we prove it for the Negative Binomial in the Supplementary Materials. The required lipschitzianity of $g^{-1}(\cdot)$ is easily met for the usual link functions (logit, identity), however, there are exceptions (log link). The modified log link function (12) in Matteson et al. (2011) provides a viable alternative. Another solution could be to replace (A6) with the alternative assumption (A3) in Douc et al. (2013), although it may be not easy to verify. Concerning the Lipschitz condition on $\tilde{g}(x)$, it depends on the shape of $\tilde{g}(x) = \bar{g}(g^{-1}(x))$, as a combination of Lipschitz function is Lipschitz continuous. A suitable choice of functions g and h will satisfy this condition. For example, when $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t^*)|\mathcal{F}_{t-1}]$, if (A5) holds, it is easy to verify that the function $\bar{g}(\mu)$ is Lipschitz with respect to (w.r.t) μ with constant not greater than 1; the same holds for g^{-1} w.r.t x, then $\tilde{g}(x)$ is Lipschitz with $L \leq 1$. When $\bar{g}(\mu_t) \neq \mathrm{E}[h(Y_t^*)|\mathcal{F}_{t-1}]$ it can be chosen accordingly to the required assumption.

Remark 2. *Let us consider equation* (3.15) *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.*

$$g(\mu_t) = \alpha + \gamma\, g(\mu_{t-1}) + \phi\, h(Y_{t-1}) + \theta\varepsilon_t, \tag{3.16}$$

where ε_t*, as in equation* (3.4)*, 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 11 still holds true by replacing* (3.15) *with* (3.16) *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*;*

The proof is deferred to the Supplementary Materials. The conditions on ν_t are widely satisfied. For example, if Y_t belongs to the exponential family in (3.3), $\sigma^2(\mu) = A^{''}(X_t) = (g^{-1})'(g(\mu))$ 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.

3.3.2 Stochastic properties for relevant encompassed models

The results obtained in the previous section can be applied to specific models belonging to the unified framework (3.2), and in particular to the novel models introduced in Section 3.2.2. We also specifically derive the stochastic properties of the related models encompassed in the framework and discussed in Section 3.2.1, since for most of them the stochastic properties have not been fully addressed in the literature. Consider the one lag models $k = p = q = 1$.

First of all, as a proof of coherence in our findings, it is worth noting that, when $\gamma = 0$ and $g \equiv h \equiv \bar{g}$, Theorem 11 reduces to Theorem 5 in Matteson et al. (2011), providing results for the GARMA model $g(\mu_t) = \alpha + \phi\, g(Y_{t-1}^*) + \theta\left[g(Y_{t-1}^*) - g(\mu_{t-1})\right]$. Now we derive the stochastic properties for the BARMA model in (3.11).

Corollary 2. *Suppose that, conditional on* $\mathcal{F}_{t-1}$*,* Y_t *is Binomial*(n, μ_t) *with fixed number of trials* n*, link function* $g : (0, a) \mapsto \mathbb{R}$ *is bijective and increasing,* g^{-1} *is Lipschitz and* $|\theta| < 1$*. Then the process* $\{\mu_t\}_{t\in T}$ *defined in* (3.11) *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 the usual link functions, like logit or probit.

At the best of our knowledge, no results are available for strict stationarity in GLARMA model, apart from the simplest case when $k = 0$, $q = 1$ (Davis et al., 2003; Dunsmuir and Scott, 2015).

Corollary 3. *Suppose that* $\{Y_t\}_{t\in T}$ *is distributed according to* (3.1). *The process* $\{\mu_t\}_{t\in T}$ *in* (3.7) *has a unique, stationary distribution and* $\{Y_t\}_{t\in T}$ *is strictly stationary and ergodic, if*

1. *g is bijective and increasing, and*

   1.1. $g : \mathbb{R} \mapsto \mathbb{R}$ *concave on* $\mathbb{R}^+$ *and convex on* $\mathbb{R}^-$, $|\gamma| < 1$

   1.2. $g : \mathbb{R}^+ \mapsto \mathbb{R}$ *concave on* $\mathbb{R}^+$, $|\gamma| + |\theta| < 1$

   1.3. $g : (0, a) \mapsto \mathbb{R}$, $|\gamma| + |\theta| < 1$.

2. g^{-1} *is Lipschitz with constant not greater than 1.*

In the GLARMA model, the conditional distribution of $\{Y_t\}_{t\in T}$ comes from the exponential family, then the (A1)-(A2) are satisfied. 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 are required to establish its stationarity.

The proof of stationarity for one lag M-GARMA model from (3.6) given in Zheng et al. (2015) only holds for continuous variable. We generalize the results 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 3.2.2.

Corollary 4. *Suppose that* $\{Y_t\}_{t\in T}$ *comes from* (3.1), $\tilde{g}(x)$ *is Lipschitz with constant* $L \leq 1$, *(A4) holds and* $|\gamma| + (|\phi| \vee |\theta|) < 1$. *Then the process* $\{\mu_t\}_{t\in T}$ *defined in* (3.12) *has a unique stationary distribution. Hence, the process* $\{Y_t\}_{t\in T}$ *is strictly stationary and ergodic.*

Assumptions (A1)-(A2) are met for the distribution (3.1). 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).

Corollary 5. *Suppose that* $\{Y_t\}_{t\in T}$ *comes from* (3.1), $\tilde{g}(x)$ *is Lipschitz with constant* $L \leq 1$ *and* $|\gamma| + |\theta| < 1$. *Then the process* $\{\mu_t\}_{t\in T}$ *defined in* (3.13) *has a unique stationary distribution. Hence, the process* $\{Y_t\}_{t\in T}$ *is strictly stationary and ergodic.*

For Binomial distribution (A1)-(A2)-(A5) hold and the conditions (A3) 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.

3.4 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 (3.1) and/or in the model (3.2), (ii) QMLE: misspecification occurs

in the distribution (3.1), (*iii*) correctly specified MLE. Specifically, strong consistency is derived in the three cases; asymptotic normality is derived for the QMLE and the correctly specified 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 Supplementary Materials.

3.4.1 Asymptotic properties

The approach of Douc et al. (2013) and Douc et al. (2017) is applied to our general framework, which is based on showing that as $t \to \infty$ the discrete-valued process $\{Y_s\}_{s\in[0,t]}$ tends to the backward infinite process $\{Y_s\}_{s\in(-\infty;t]}$, the latter is then used to establish the asymptotic properties of the likelihood estimator. See the Appendix for details. Assume that $\{Y_n\}_{n\in\mathbb{Z}}$ are integer-valued. Let (Λ, d) be a compact metric set of parameter, with suitable metric $d(\cdot)$, and $\Lambda = \left\{\rho = (\alpha, \gamma, \phi, \theta) \in \mathbb{R}^4 : |\alpha| \leq \tilde{\alpha}, |\delta| = |\phi + \theta| \leq \tilde{\delta}\right\}$, where $\tilde{\alpha}, \tilde{\delta} \in \mathbb{R}^+$. We make explicit the dependence of the conditional distribution (3.1) 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 (3.15), such that

$$g^\rho\langle Y_{-\infty:t}\rangle = \alpha + \gamma g^\rho\langle Y_{-\infty:t-1}\rangle + \phi h(Y_{t-1}) + \theta[h(Y_{t-1}) - \tilde{g}(g^\rho\langle Y_{-\infty:t-1}\rangle)]\,, \tag{3.17}$$

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 Diaconis and Freedman (1999), with

$$g^\rho_{y_1}(x) = \alpha + \gamma x + \phi h(y_0) + \theta[h(y_0) - \tilde{g}(x)]\,. \tag{3.18}$$

It is worth noting that in the special case of correctly specified model, $X_0 = g^\rho\langle Y_{-\infty:0}\rangle$ and equation (3.17) reduces exactly to the process in equation (3.15). 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

$$\hat{\rho}_{n,x} = \arg\max_{\rho\in\Lambda} L^\rho_{n,x}\langle Y_{1:n}\rangle\,. \tag{3.19}$$

Consider 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 in the Supplementary Materials 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 12. *Assume that Theorem 11 and* (H1) *hold. Then,* $\forall x \in S$, $\lim_{n\to\infty} d(\hat{\rho}_{n,x}, \mathrm{P}_\star) = 0$, *a.s., where* $\mathrm{P}_\star \coloneqq \arg\max_{\rho\in\Lambda} \mathrm{E}\{Y_0\, f[g^\rho\langle Y_{-\infty:0}\rangle] - A[g^\rho\langle Y_{-\infty:0}\rangle] + d(Y_0)\}$.

Here, the almost sure limit is meant to be valid under the stationary distribution of $\{Y_t\}_{t\in T}$. The proof lies in the Appendix. Now the special case of correctly specified MLE is treated.

Theorem 13. *Assume that* $\{Y_n\}_{n\in\mathbb{Z}}$ *is distributed according to* (3.1) *and satisfies the recursion* (3.15)*, with parameters* $\rho_\star \in \Lambda^0$. *Moreover, assume that Theorem 12 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. Let us denote Λ^0 as the interior of the set Λ.

Corollary 6. *Assume that $\{Y_n\}_{n\in\mathbb{Z}}$ satisfies the recursion (3.15), with parameters $\rho_\star \in \Lambda^0$ and $\mu = A'(x_\star)$. Moreover, assume that Theorem 12 holds. Then, for all $x \in S$,*

$$\lim_{n\to\infty} \hat{\rho}_{n,x} = \rho_\star, \quad a.s. \tag{3.20}$$

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 13 and follows directly by Theorem 4.1 and Douc et al. (2017, Thr 4.1). Finally, we investigate the conditions under which the QMLE (3.20) is asymptotically normally distributed for the model (3.15).

Theorem 14. *Assume that Corollary 6 and (H2) hold. Moreover, assume that the matrix (3.21) 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*

$$\begin{aligned}
\mathcal{I}(\rho_\star) &\coloneqq \mathrm{E}\left[\left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right) \left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right)' \left(\frac{\partial}{\partial x} \log q\left(g^{\rho_\star}\langle Y_{-\infty:0}\rangle, Y_1\right)\right)^2 \right], \\
\mathcal{J}(\rho_\star) &\coloneqq \mathrm{E}\left[\left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right) \left(\nabla_\rho g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right)' \frac{\partial^2}{\partial x^2} \log q\left(g^{\rho_\star}\langle Y_{-\infty:0}\rangle, Y_1\right)\right].
\end{aligned} \tag{3.21}$$

The proof relies on the argument of Douc et al. (2017, Thr 4.2) and follows the fashion and the notation used in the proof of Theorem 12, thus it is postponed to the Supplementary Materials. It goes without saying that for correctly specified MLE, equation (3.19) is the exact MLE and $\mathcal{J}(\rho_\star) = \mathcal{I}(\rho_\star)$ in Theorem 14, providing the standard ML inference.

3.4.2 Finite sample properties and model selection

Finite sample properties of MLE and QMLE are explored through a simulation study which considers some models illustrated in Sections 3.2.1 and 3.2.2. The details of the numerical results are stored in the Supplementary Materials. All the results are based on $s = 1000$ replications, with different configuration of the parameters and increasing sample size $n = (200, 500, 1000, 2000)$. A correctly specified MLE has been carried out with data coming from Bernoulli or Poisson distributions across several models. Simulations of QMLE are performed on data generated from Geometric distribution, with Poisson distribution fitted instead, for GARMA and log-AR model. 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 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 emerge from the Kolmogorov-Smirnov test, even when the sample size is small. The outcomes are in line with those of Douc et al. (2017). These results are coherent with the theory presented so far.

A crucial aspect in empirical applications is model selection. In likelihood inference, model selection is typically carried out based on information criteria such as the Akaike information criterion (AIC) and the 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. The details of the analysis are reported in the Supplementary Materials. To summarize the results, when the sample size n is small, the selection for some models can perform poorly, but when n is big enough, all the models allow to select the right data generating model with high probability.

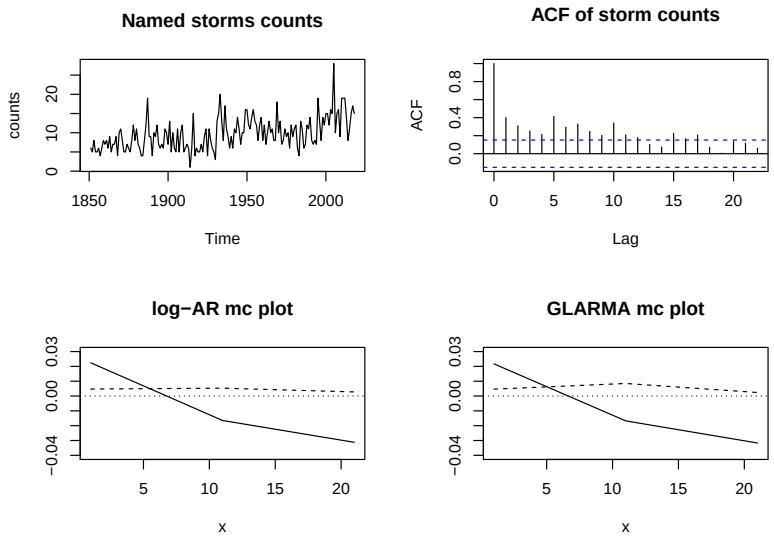

Figure 3.1: Top-left: storms counts. Top-right: ACF. Bottom-right: mc plot for GLARMA model. Bottom-left: mc plot for log-AR model. Dashed line is Poisson. Black line NB.

3.5 Applications

3.5.1 Number of storms in the North Atlantic Basin

We apply the dynamic models discussed so far for a novel application based on a set of data related to the annual number of named storms in the North Atlantic Hurricane Basin from 1851 to 2018; counts of storms are related to tropical storms, hurricanes and subtropical storms. The data can be found in the revised HURDAT database at `https://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html`. There is an intense scientific debate over the increasing hurricane activity to figure out whether hurricanes are becoming more numerous, or whether the strengths of storms are increasing, mainly because of the warming earth. Then the prediction of the number of storms is crucial and becomes of primary interest; see Villarini et al. (2010) for a discussion and Livsey et al. (2018) of a recent application in a similar context. The time series is relatively short $n = 168$ and is plotted in Figure 3.1 along with the sample autocorrelation function (ACF). There is a temporal correlation which spreads over several lags. For the data generating process we assume both the Poisson and the Negative Binomial (NB) distribution in equation (3.10), where $\nu > 0$ is the dispersion parameter and μ_t is the conditional expectation; the latter is the same for both distributions. Indeed, equation (3.10) is defined in terms of mean rather than of the probability parameter $p_t = \frac{\nu}{\nu+\mu_t}$ and, unlike the case of Poisson distribution, 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 some models belonging to the class in equation (3.15):

$$\text{log-AR:} \qquad \log(\mu_t) = \alpha + \phi \log(y_{t-1} + 1) + \gamma \log(\mu_{t-1})\,,$$

$$\text{GARMA:} \qquad \log(\mu_t) = \alpha + \phi \log(y^\star_{t-1}) + \theta \left[\log(y^\star_{t-1}) - \log(\mu_{t-1})\right]\,,$$

$$\text{GLARMA:} \qquad \log(\mu_t) = \alpha + \gamma \log(\mu_{t-1}) + \theta \left(\tfrac{y_{t-1}-\mu_{t-1}}{s_{t-1}}\right)\,,$$

where $y^\star_t = \max\{y_t, c\}$ with $c = 0.1$. Different values of $0 < c < 1$ did 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

Table 3.1: MLE results for named storms.

Models	$\hat{\alpha}$	$\hat{\phi}$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\nu}$	*AIC*	*BIC*	*QIC*
Pois log-AR	0.212	0.231	0.673	-	-	11.361	20.733	8.881
	(0.082)	(0.058)	(0.089)	-				
Pois GARMA	0.289	0.882	-	-0.684	-	11.368	20.740	8.644
	(0.092)	(0.039)	-	(0.083)				
Pois GLARMA	0.314	-	0.864	0.071	-	**11.359**	**20.731**	9.187
	(0.103)	-	(0.046)	(0.018)				
NB log-AR	0.390	0.286	0.540	-	5.262	11.528	20.900	8.810
	(0.310)	(0.114)	(0.246)	-				
NB GARMA	0.483	0.797	-	-0.556	5.190	11.536	20.908	8.913
	(0.354)	(0.154)	-	(0.248)				
NB GLARMA	0.376	-	0.836	0.139	5.402	11.510	20.881	**7.640**
	(0.194)	-	(0.086)	(0.041)				

Information Criterion (QIC) introduced by Pan (2001) 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 case of well-specified models. QMLE estimation 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 BFGS algorithm. The score functions written in terms of predictor $x_t = \log \mu_t$ are:

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

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 Christou and Fokianos (2015) is used:

$$\hat{\nu} = \left(\frac{1}{n}\sum_{t=1}^{n}\frac{(y_t-\hat{\mu}_t)^2 - \hat{\mu}_t}{\hat{\mu}_t^2}\right)^{-1} \tag{3.22}$$

where $\hat{\mu}_t = \mu_t(\hat{\rho})$ comes 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 (3.22), 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 14; each quantity has been replaced by its sample counterpart.

The results related to MLEs are summarized in Table 3.1. The intercept is not significant, at 5% level, for the NB log-AR and GARMA models. All the other coefficients are significant. The parameter $\hat{\nu}$ is generally around 5. Both AIC and BIC select the Pois GLARMA model as the best, in a goodness-of-fit sense, followed by the Pois Log-AR. The QIC selects the GLARMA model, as well, but with NB distribution. This might be an indication of overdispersion in the true data generating process, not captured by the Poisson models; this hypothesis is also supported and discussed in Villarini et al. (2010).

We then assess the adequacy of the fit. We check 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 by the ACF, which in our case appears uncorrelated. Another check comes from the probability and

marginal calibrations, as defined in Gneiting et al. (2007). Czado et al. (2009) introduced a non-randomized version of Probability Integral Transform (PIT) for discrete data. It can be built based on the conditional cumulative distribution function

$$F(u|y_t) = \begin{cases} 0, & u \leq P_t(y_t - 1) \\ \frac{u - P_t(y_t-1)}{P_t(y_t) - P_t(y_t-1)}, & P_t(y_t) \leq u \leq P_t(y_t - 1) \\ 1, & u \geq P_t(y_t) \end{cases} \tag{3.23}$$

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 (3.23) will appear to be the cumulative distribution function of a Uniform(0,1). The PIT (3.23) is computed for each realisation 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 $\bar{F}(\frac{j}{J}) - \bar{F}(\frac{j-1}{J})$; a representative plot is in the Supplementary Material, Figure S-1. The difference between the distributions is subtle but the Poisson PIT's seems to be closer to Uniform(0,1). The marginal calibration (mc) is assessed as in Gneiting et al. (2007) and Christou and Fokianos (2015). 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, $\bar{G}(x) = 1/n \sum_{t=1}^{n} \mathbf{1}(y_t \leq x)$. A plot of the outcomes for mc is in Figure 3.1 for log-AR and GLARMA model. In the other models the results are similar. It appears a better concordance with empirical distribution for the Poisson case.

In order to assess the predictive power, we refer to the concept of sharpness of the predictive distribution defined in Gneiting et al. (2007). It can be measured by some average quantities related to the predictive distribution, which take 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) $\sum_{k=0}^{\infty}[P_t(k) - \mathbf{1}(y_t \leq k)]$, for different models and distributions. Then, the predictive performance is evaluated and the Poisson log-AR model provides the best predictive performance for 3 up to 4 scoring rules. Numerical results for each model are collected in Table S-5 in the Supplementary Material. This leads to a different model selection, depending on the aims of the empirical analysis.

3.5.2 Disease cases of Escherichia coli in North Rhine-Westphalia

We consider a testbed set of data related to 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`. The time series has a time length $n = 646$ and is plotted in Figure 3.2, with its sample ACF. There is a temporal correlation which spreads over several lags with a greater magnitude compared to the dataset in the previous example. The slow decay of the ACF suggests the use of a feedback mechanism. The same models, distributions and estimation procedures of the storm application have been employed.

The results of the analysis are summarized in Table 3.2. 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. The ACF of the residuals appears uncorrelated. A plot representative of PIT value is in the Supplementary Material, Figure S-2. The NB seems to be more appropriate for our data as its PIT's are quite near to Uniform(0,1). The marginal calibration (mc) is plotted in Figure 3.2 for log-AR and GLARMA model. 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. Results related to the predictive power are summarized in Table S-6 in the Supplementary Material. The NB GLARMA model has the best predictive performance for the majority of the scores and it is ultimately chosen since it has been also selected by the information criteria.

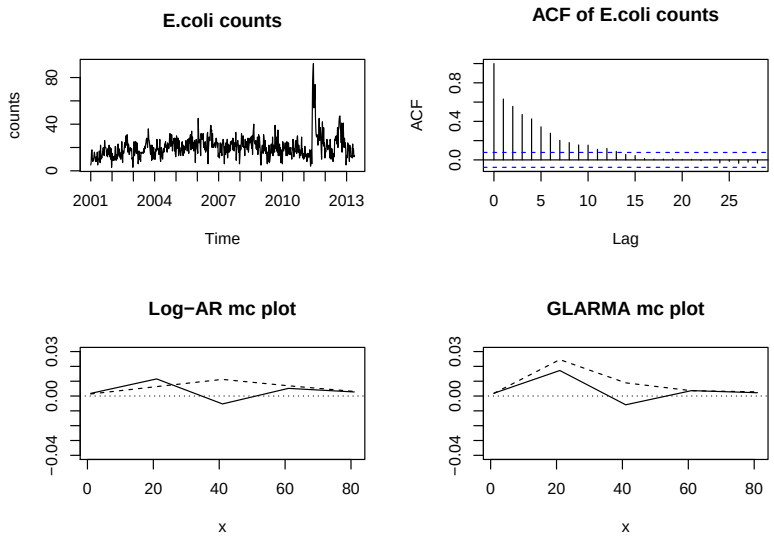

Figure 3.2: Top-left: Escherichia coli counts. Top-right: ACF. Bottom-left: mc plot log-AR. Bottom-right: mc plot for GLARMA model. Dashed line is Poisson. Black line is NB.

3.6 Discussion

We developed statistical inference for a class of observation driven models which encompasses known models as well as new models of potential interest for the analysis of discrete-valued time series. Strict stationarity and ergodicity conditions have been derived for any model in the class and a large family of probability distributions satisfying mild moment conditions. Consistency and asymptotic normality of the quasi maximum likelihood estimators have been also established, with the focus on the exponential family. We expect the specification of this broad class of models will provide useful theoretical and modelling enhancements to study the dynamic trend of count and binary data.

From a theoretical perspective, the unified framework permits to generalize the results on stochastic and inferential properties for well-known models and to establish the same results for new models introduced in Section 3.2.2 of potential interest. Although the uniqueness of the stationary distribution for the process is proved in Section 3.3 by using Markov chain theory, the rate of convergence to the limiting distribution still represents an open issue. Improvements could be achieved by considering a Markov chain of order greater than 1 to define a model with several lags besides the first.

From a modelling side, the proposed framework allows one to accounts for three relevant aspects in the analysis of temporal data: (i) the autoregressive-like effect, (ii) the effect of the past memory dependence and (iii) the effect of the moving average part. Models in the class may differ for the effects they consider and also for the specification of these effects through suitable link-functions. Then, the merit of the unified framework is to provide a wide range of dynamic models which could be extremely different, not necessarily nested, but fully applicable and comparable in practice since they belong to the same class. Model selection in terms of fitting and prediction across different models can be performed using information criteria; their performance is explored through an extensive simulation study.

Finally, in line with the recent theory developed for some multivariate discrete-valued processes (Fokianos et al., 2020), the specification of a unified framework for modelling multivariate discrete-valued time series may represent an interesting generalization.

Table 3.2: 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	12.197
	(0.102)	(0.05)	(0.073)	-				
NB GARMA	0.640	0.794	-	-0.420	9.865	12.641	26.053	12.336
	(0.111)	(0.036)	-	(0.074)				
NB GLARMA	0.483	-	0.839	0.142	10.892	**12.578**	**25.990**	**11.895**
	(0.110)	-	(0.036)	(0.019)				

Appendix

Proof of Theorem 12

Proof. Equation (3.18) 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 $s = 0, \dots, t$ 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

$$g^{\rho}_{y_1}(x) = \alpha + \gamma x + (\phi + \theta)h(y_0) - \theta\tilde{g}(x) = \alpha + \delta h(y_0) + r_0 x \tag{B-1}$$

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

$$g^{\rho}\langle y_{s:t}\rangle(x) = \alpha \sum_{j=0}^{t-s} \prod_{i=0}^{j-1} r_{t-i} + \delta \sum_{j=0}^{t-s} \prod_{i=0}^{j-1} r_{t-i} h(y^*_{t-j}) + \prod_{j=0}^{t-s} r_j x\,, \tag{B-2}$$

where $r_{t-i} = 1$ for $i = -1$. Moreover, from (B-2), and by equation (3.17), 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 Lipschitzianity of $\tilde{g}$. Then, from Theorem 11, 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 (B-1) also holds in the simpler case $\bar{g}(\cdot) = g(\cdot)$, with $r_0 = r = \gamma - \theta$, where $|\gamma - \theta| < 1$ from Theorem 11. Let $\{Y_n\}_{n\in\mathbb{Z}}$ be a strictly stationary and ergodic process, satisfying Theorem 11. The proof of Theorem 12 holds if assumptions (B1)-(B3) in Douc et al. (Thr. 19, 2013) 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 Douc et al. (Rmk. 18, 2013). 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.

$$\sup_{\rho\in\Lambda}|g^{\rho}\langle Y_{-\infty:0}\rangle| \leq |\alpha|\sum_{j=0}^{\infty}\tilde{r}^{j} + |\delta|\sum_{j=0}^{\infty}\tilde{r}^{j}|h(Y^{*}_{-j})| \leq \frac{\tilde{\alpha}}{1-\tilde{r}} + \tilde{\delta}\sum_{j=0}^{\infty}\tilde{r}^{j}|h(Y^{*}_{-j})| = \hat{g}\langle Y_{-\infty:0}\rangle\,, \tag{B-3}$$

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 $\mathrm{E}[Y^{*}_{0}] \leq \mathrm{E}[Y_0] + c$ (see equation (S-8) in the Supplementary Materials). In Case 3 $h(\cdot)$ and Y_t are bounded so their expectations are finite. It holds also that

$$|g^{\rho}\langle Y_{-\infty:t-1}\rangle| \leq \frac{\tilde{\alpha}}{1-\tilde{r}} + \tilde{\delta}\sum_{j=0}^{\infty}\tilde{r}^{j}|h(Y^{*}_{t-1-j})| \tag{B-4}$$

$$|g^{\rho}\langle Y_{1:t-1}\rangle(x)| \leq \tilde{\alpha}\sum_{j=0}^{t-2}\tilde{r}^{j} + \tilde{\delta}\sum_{j=0}^{t-2}\tilde{r}^{j}|h(Y^{*}_{t-1-j})| + \tilde{r}^{t-1}|x| \tag{B-5}$$

which possesses a finite expectation according to (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,

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

converges to 0 as $m \to \infty$ by (H1) and Douc et al. (2013, Lem. 34). Thus (B3)-(i) holds. We now move to (B3)-(ii),

$$\begin{aligned}
&\sup_{\rho\in\Lambda}|\log q(g^{\rho}\langle Y_{1:t-1}\rangle(x); Y_t) - \log q(g^{\rho}\langle Y_{-\infty:t-1}\rangle; Y_t)| \\
&\leq Y_t \sup_{\rho\in\Lambda}|f[g^{\rho}\langle Y_{1:t-1}\rangle(x)] - f[g^{\rho}\langle Y_{-\infty:t-1}\rangle]\,| + \sup_{\rho\in\Lambda}|A\left[g^{\rho}\langle Y_{1:t-1}\rangle(x)\right] - A\left[g^{\rho}\langle Y_{-\infty:t-1}\rangle\right]|\,.
\end{aligned}$$

First consider

$$\begin{aligned}
|g^{\rho}\langle Y_{1:t-1}\rangle(x) - g^{\rho}\langle Y_{-\infty:t-1}\rangle| \quad &= \left|\alpha\sum_{l=0}^{\infty}\prod_{i=0}^{t+l-2} r_{t-1-i} + \delta\sum_{l=0}^{\infty}\prod_{i=0}^{t+l-2} r_{t-1-i}h(Y^{*}_{-l}) + \prod_{j=0}^{t-2} r_j x\right| \\
&\leq \tilde{r}^{t-1}\left(\tilde{\alpha}\sum_{l=0}^{\infty}\tilde{r}^{l} + \tilde{\delta}\sum_{l=0}^{\infty}\tilde{r}^{l-}|h(Y^{*}_{-l})| + |x|\right) \\
&= \tilde{r}^{t-1}\left(|x| + \hat{g}\langle Y_{-\infty:0}\rangle\right)
\end{aligned}$$

for (B-3), and for $l = j$ when $t - 1 = 0$. This implies that

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

according to (B-3) and by Douc et al. (2013, Lem. 34), under (H1). Now, for the mean value theorem,

$$\begin{aligned}
\sup_{\rho\in\Lambda}|A\left[g^{\rho}\langle Y_{1:t-1}\rangle(x)\right] - A\left[g^{\rho}\langle Y_{-\infty:t-1}\rangle\right]| \quad &= \sup_{\rho\in\Lambda}|A'(C_{t-1})|\,|g^{\rho}\langle Y_{1:t-1}\rangle(x) - g^{\rho}\langle Y_{-\infty:t-1}\rangle| \\
&\leq \sup_{\rho\in\Lambda}|A'(C_{t-1})|\,\tilde{r}^{t-1}\left(|x| + \hat{g}\langle Y_{-\infty:0}\rangle\right)
\end{aligned} \tag{B-6}$$

where $\min\{g^{\rho}\langle Y_{1:t-1}\rangle(x), g^{\rho}\langle Y_{-\infty:t-1}\rangle\} \leq C_{t-1} \leq \max\{g^{\rho}\langle Y_{1:t-1}\rangle(x), g^{\rho}\langle Y_{-\infty:t-1}\rangle\}$. The function (B-6) tends to 0 as $t \to \infty$, for Douc et al. (2013, Lem. 34) and $\mathrm{E}[(\log|A'(C_{t-1})|)_+] < \infty$, which is true for (H1). The same argument of (B-6) hold with $f(\cdot)$ instead of $A(\cdot)$, and the details are omitted. Then, (B3)-(ii) holds, and this completes the proof. □

Proof of Theorem 13

Proof. First of all, we note that $P(x, A) = \int_A q(x; y)\mu(dy)$. By the stationarity of Y_t and (H1), Theorem 12 holds. It remains to show that $\mathrm{P}_\star = \{\rho_\star\}$, where $\rho_\star = (\alpha_\star, \gamma_\star, \phi_\star, \theta_\star)$. This follows from Douc et al. (Prop. 21, 2013), once we have showed that

(LP1) $X_0 = g^{\rho_\star}\langle Y_{-\infty:0}\rangle$, a.s.

(LP2) $x \mapsto P(x; \cdot)$ is one-to-one mapping, 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 $\rho = \rho_\star$.

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

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

where $\delta_\star = \phi_\star + \theta_\star$, $r_{\star s} = \gamma_\star - \theta_\star c_s$ for $-j+1 \leq s \leq 0$. 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. □

Supplementary Material

This is a supplementary material containing proofs of Theorem 11, Theorem 14 and Corollary 1. The equivalence of (A4) and (A5) for the Negative Binomial is verified. Some insight about conditions (H1)-(H2) is provided. Moreover, the numerical results of the simulation study discussed in Section 3.4.2 are reported. Finally, additional numerical results for the application in Section 3.5 are showed.

Main proofs

Preliminary Lemmata for Proof of Theorem 11

The proof of Theorem 11 requires some definitions and preliminary lemmata, with the same notation of Theorem 11.

Definition 9. *A set $A \in \mathcal{F}$ is called a small set if there exists $m > 1$, a nontrivial measure v on $\mathcal{F}$, and $\lambda > 0$ such that $\forall x \in A$, $\forall C \in \mathcal{F}$, $P^m(x, C) \geq \lambda\, v(C)$.*

Definition 10. *A chain evolving on a complete separable metric space S is said to be "weak Feller" if $P(x, \cdot)$ satisfies $P(x, \cdot) \Rightarrow P(y, \cdot)$ as $x \to y$, for any $y \in S$ and where $\Rightarrow$ indicates convergence in distribution.*

Definition 11. *Let S be a Polish (complete, separable, metrizable) space. A "totally separating system of metrics" $\{d_t\}_{t\in N}$ for S is a set of metrics such that for any $x, y \in S$ with $x \neq y$, the value $d_t(x, y)$ is nondecreasing in t and $\lim_{t\to\infty} d_t(x, y) = 1$.*

Definition 12. *A chain is "asymptotically strong Feller" if, for every fixed $x \in S$, there is a totally separating system of metric $\{d_t\}$ for S and a sequence $t_n > 0$ such that*

$$\lim_{\delta\to\infty} \limsup_{t\to\infty} \sup_{y\in B(x,\delta)} \left\| P^{t_n}(x, \cdot) - P^{t_n}(y, \cdot) \right\|_{d_t} = 0$$

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

Definition 13. *A "reachable" point $x \in S$ means that $\forall$ open sets A containing x, $\forall y \in S$, $\sum_{t=1}^{\infty} P^t(y, A) > 0$.*

The proof of Theorem 11 is essentially based on the following preliminary lemmata. First, a drift condition is proven on the Markov chain X_t (Lemma 3); after that, the weak Feller property is established for the chain (Lemma 4), which proves the existence of a stationary distribution for $\{X_t\}_{t\in T}$. Then, the asymptotic strong Feller condition is verified (Lemma 5). Finally, the existence of a reachable point is shown (Lemma 6) and, by combining all these results, the uniqueness of the stationary distribution of the chain is proven.

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 $\forall x \in S$:

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

where $\eta \in (0, 1)$, $b > 0$, $V : S \to [1, \infty)$ and $A \subset S$ is a small set.
Let (A3.1) g and h are bijective, increasing and

1. If $\bar{g}(\mu_t) = g(\mu_t)$,
   - 1.1. $h : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $g : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $|\gamma| + |\phi| < 1$
   - 1.2. $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $(|\gamma| + |\phi|) \vee |\gamma - \theta| < 1$
   - 1.3. $h : (0, a) \mapsto \mathbb{R}$ and $g : (0, a) \mapsto \mathbb{R}$, $|\gamma - \theta| < 1$.
2. If $\bar{g}(\mu_t) \neq g(\mu_t)$ and $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t^*)|\mathcal{F}_{t-1}]$ or $\bar{g} \equiv h$
   - 2.1. $h : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $g : \mathbb{R} \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$ and convex on $\mathbb{R}^-$, $|\gamma| + |\phi| < 1$
   - 2.2. $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on $\mathbb{R}^+$, $|\gamma| + |\phi| < 1$
   - 2.3. $h : (0, a) \mapsto \mathbb{R}$ and $g : (0, a) \mapsto \mathbb{R}$, $|\gamma| < 1$.

Lemma 3. *Under assumptions* (A1), (A2) *and* (A3.1), *the chain $\{X_t\}_{t\in T}$ has a small set $A \subset S$ and satisfies the drift condition* (S-1).

Proof of Lemma 3

Proof. The proof is inspired on Matteson et al. (Sec. 4.1, 2011) and the propositions therein. Firstly, we define a small set $A = [-M, M]$ for some constant $M > 0$, where it is known 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. For details see Matteson et al. (p. 812, 2011). Next, it is possible to use the small set A to prove the drift condition (S-1) by taking the function $V(x) = |x|$. Then, we split the drift condition in three parts of the real axis: $x < -M, \quad x \in [-M, M], \quad x > M$. Only the parts of the proof which differ significantly are shown. First we will give the drift condition for $x \in A$,

Proposition 2. ***(Cases 1-3)*** *There is some constant $G(M) < \infty$ such that $\mathrm{E}_x V(X_1) \leq G(M)$ for all $x \in A$.*

Then, the drift condition for $x \notin A$ is provided, handling the cases $x < -M$ and $x > M$ separately.

Proposition 3. ***(Case 1)*** *For any $\varepsilon \in (0,1)$ there is some constant $G_2 < \infty$ such that for M large enough, $\mathrm{E}_x V(X_1) \leq (|\phi| + |\gamma| + \varepsilon)V(x) + G_2$ for all $x < -M$.*
(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 $x < -M$.*

Proposition 4. ***(Cases 1-2)*** *For any $\varepsilon \in (0,1)$ there is some constant $G_3 < \infty$ such that for M large enough, $\mathrm{E}_x V(X_1) \leq (|\phi| + |\gamma| + \varepsilon)V(x) + G_3$ for all $x > M$.*
(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 $x > M$.*

Propositions 3 and 4 give the overall drift condition for $x \notin A$ as follows. Consider Case 2; the other two cases are analogous. If $\bar{g}(\mu) = g(\mu)$, since $\varepsilon > 0$, we can write $\mathrm{E}_x V(X_1) \leq |\gamma - \theta|V(x) + W_2 \leq (|\gamma - \theta| + \varepsilon)V(x) + W_2$ from Proposition 3 and, for M large enough, $\mathrm{E}_x V(X_1) \leq (|\phi| + |\gamma| + \varepsilon)V(x) + G_3$ from Proposition 4. 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 to satisfy Proposition 4. Then, for any $x \notin A$, we have $\mathrm{E}_x V(X_1) \leq \eta V(x) + L$, establishing the drift condition (S-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 $\eta = |\phi| + |\gamma| + \varepsilon$, establishing the drift condition (S-1) for $|\phi| + |\gamma| < 1$.

Proof of Proposition 2, Case 1

We assume, without loss of generality (w.l.o.g.), 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; same assumptions hold for g. Now, we can bound $\mathrm{E}_x V(X_1)$ to obtain the drift condition (S-1) as follows, where C, C_1, C_2 denote bounded constants with respect to μ which can take different values:

$$\begin{aligned} \mathrm{E}_x V(X_1) &= \mathrm{E}_x|\alpha + \gamma x + \phi h(Y_0) + \theta[h(Y_0) - \bar{g}(\mu)]| \\ &\leq |\alpha| + |\gamma|\mathrm{E}_x|x| + |\phi|\mathrm{E}_x|h(Y_0)| + |\theta|\mathrm{E}_x|h(Y_0) - \bar{g}(\mu)| \\ &\leq |\alpha| + |\gamma||x| + (|\phi| + |\theta|)a_1\mathrm{E}_x|Y_0| + |\theta||\tilde{g}(x)|\,. \end{aligned} \tag{S-2}$$

From Matteson et al. (p. 21, 2011), $\mathrm{E}_x|Y_0|$ is bounded. So $\sup_{x \in [-M,M]} \mathrm{E}_x V(X_1) < \infty$, proving Proposition 2.

Proof of Proposition 3 and 4, Case 1

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

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

When $h(\mu) \leq g(\mu)$, this holds from a result in Matteson et al. (Sec. A.7, 2011) by substituting $h(\cdot)$ to $g(\cdot)$. Instead, when $h(\mu) > g(\mu)$, the results 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 need to show that the term $\mathrm{E}_x|h(Y_0) - \bar{g}(\mu)|$ in (S-2) is "small" relative to the linear term in x:

Proposition 5. *There are some constants l_{14}, l_{15} 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.*

Proof of Proposition 5

Since $h(0) = 0$ and h is monotonic increasing, for $x > M$, by Matteson et al. (eq. 23, 2011),

$$\begin{aligned}
\mathrm{E}_x|h(Y_0) - \bar{g}(\mu)| &= \mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu) + h(Y_0\mathbf{1}_{Y_0<0})| \\
&\leq \mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu)| + \mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0<0})| \\
&\leq \mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu)| + a_0 + a_1\mathrm{E}_x[|Y_0|\mathbf{1}_{Y_0<0}] \\
&\leq \mathrm{E}_x|h(Y_0\mathbf{1}_{Y_0>0}) - \bar{g}(\mu)| + C.
\end{aligned}$$

Using the Markov inequality stated in Matteson et al. (eq. 14, 2011), for any fixed $\varepsilon \in (0,1)$ and $x > M$,

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

If $\bar{g} \equiv h \neq g$, equation (S-5) reduces to $\frac{Ch(\mu)}{\mu^{2+\delta-r}}$. 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 (S-4) is bounded by: $\frac{Ch(\mu)}{[h(\mu)-a_0]^{2+\delta-r}} = \frac{C\tilde{h}(x)}{[\tilde{h}(x)-a_0]^{2+\delta-r}}$ 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 (S-5) as $\frac{g(\mu)(C_1\mu^r+C_2)}{\varepsilon^{2+\delta}\mu^{2+\delta}} + \frac{h(\mu)(C_1\mu^r+C_2)}{\varepsilon^{2+\delta}\mu^{2+\delta}} = \frac{Cx}{\mu^{2+\delta-r}} + \frac{Ch(\mu)}{\mu^{2+\delta-r}}$ and it is bounded by $\frac{Cx}{[x-a_0]^{2+\delta-r}} + \frac{Ch(\mu)}{[h(\mu)-a_0]^{2+\delta-r}} = \frac{Cx}{[x-a_0]^{2+\delta-r}} + \frac{C\tilde{h}(x)}{[\tilde{h}(x)-a_0]^{2+\delta-r}}$, which converges to 0 as $x \to \infty$. It only remains to show that

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

is "small". When $\bar{g} \equiv h$, this is straightforward by substituting $h(\cdot)$ to $g(\cdot)$ in Matteson et al. (p. 826, 2011), establishing Proposition 5. For $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0)]$, the expectation (S-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 $x > M$, (p. 824, Matteson et al., 2011), since $\mu \leq \frac{x}{b_1(x)(1-\varepsilon)}$ by equation (S-1) where $b_1(x)$ is bounded for $x > M$. Then, Proposition 5 is proved also for $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0)]$.

Combining Proposition 5 with (S-2) and (S-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 proves Proposition 4. Proposition 3 holds by symmetry when $x < -M$.

Proof of Proposition 2, 3, Case 2

Assume w.l.o.g. that $h(c) = 0$ and $g(c) = 0$, this simply changes the value of α. Since $h(c) = 0$, $h(Y_0^*) \geq 0$ is non-negative for any Y_0^*. Also, due to the concavity of h, there is some $a_1 > 0$ such that $h(y) \leq a_1 y$ for all $y \in \mathbb{R}^+$. The same arguments hold for g. At this point, a different proof is developed, depending on the shape of $\bar{g}(\mu)$. When $\tilde{g}(x) = \bar{g}(\mu)$, we can bound $\mathrm{E}_x V(X_1)$ as follows:

$$\begin{aligned}
\mathrm{E}_x V(X_1) &= \mathrm{E}_x|\alpha + \gamma x + \phi h(Y_0^*) + \theta[h(Y_0^*) - \bar{g}(\mu)]| \\
&\leq |\alpha| + |\gamma - \theta||x| + |\phi + \theta|\mathrm{E}_x[h(Y_0^*)] \\
&= |\alpha| + |\phi + \theta|\mathrm{P}_x(Y_0 < c)h(c) + |\phi + \theta|\mathrm{E}_x[h(Y_0)\mathbf{1}_{Y_0 \geq c}] + |\gamma - \theta||x| \\
&= |\alpha| + |\phi + \theta| a_1 \mathrm{E}_x[Y_0 \mathbf{1}_{Y_0 \geq c}] + |\gamma - \theta||x|\,.
\end{aligned}$$

Note that $\mathrm{E}_x[Y_0 \mathbf{1}_{Y_0 \geq c}] \leq \mathrm{E}_x|Y_0| \leq C_2 + C_1 \mu^{r/(2+\delta)}$ where $\mu = g^{-1}(x)$, implying that $\mathrm{E}_x V(X_1) \leq C_2 + C_1\mu + |\theta||\tilde{g}(x)| + |\gamma||x|\,$, so $\mathrm{E}_x V(X_1) < \infty$ for $x \in [-M, M]$, proving Proposition 2. When $x < -M$ we have $\mu = g^{-1}(x) \leq g^{-1}(0) = c$, we obtain $\mathrm{E}_x V(X_1) \leq l_{20} + |\gamma - \theta||x|$, and this completes the proof of Proposition 3.

Now the case when $\bar{g} \neq g$ is considered. A different bound for $\mathrm{E}_x V(X_1)$ can be established:

$$\begin{aligned}
\mathrm{E}_x V(X_1) &= \mathrm{E}_x|\alpha + \gamma x + \phi h(Y_0^*) + \theta[h(Y_0^*) - \bar{g}(\mu)]| \\
&\leq |\alpha| + |\gamma||x| + |\phi|\mathrm{E}_x[h(Y_0^*)] + |\theta|\mathrm{E}_x|h(Y_0^*) - \bar{g}(\mu)| \\
&= |\alpha| + |\phi|\mathrm{P}_x(Y_0 < c)h(c) + |\phi|\mathrm{E}_x[h(Y_0)\mathbf{1}_{Y_0 \geq c}] + |\gamma||x| \\
&\quad + |\theta|\mathrm{P}_x(Y_0 < c)|h(c) - \bar{g}(\mu)| + |\theta|\mathrm{E}_x[|h(Y_0) - \bar{g}(\mu)|\mathbf{1}_{Y_0 \geq c}] \\
&\leq |\alpha| + (|\phi| + |\theta|)\mathrm{E}_x[h(Y_0)\mathbf{1}_{Y_0 \geq c}] + |\theta|\mathrm{P}_x(Y_0 < c)|h(c) - \bar{g}(\mu)| \\
&\quad + |\theta|\mathrm{P}_x(Y_0 \geq c)|\bar{g}(\mu)| + |\gamma||x| \\
&\leq |\alpha| + (|\phi| + |\theta|)a_1\mathrm{E}_x[Y_0\mathbf{1}_{Y_0 \geq c}] + |\theta||\tilde{g}(x)| + |\gamma||x|\,.
\end{aligned} \tag{S-7}$$

When $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0^*)]$, $\mathrm{E}_x[h(Y_0^*)] = \mathrm{E}_x[h(Y_0\mathbf{1}_{Y_0 \geq c}) + h(c)\mathrm{P}_x(Y_0 < c)] \leq a_1[\mathrm{E}_x(Y_0\mathbf{1}_{Y_0 \geq c})] \leq a_1[\mathrm{E}_x|Y_0|] \leq C_1 + C_2\mu^{r/(2+\delta)}$ and so $\mathrm{E}_x V(X_1) \leq C + |\theta||\mathrm{E}_x[h(Y_0^*)]| + |\gamma||x| \leq C + |\theta|(C_2 + C_1 c^{r/(2+\delta)}) + |\gamma||x|\,$. Lastly, if $\bar{g} \equiv h$, $|\tilde{g}(x)| = |h(\mu)| = -h(\mu) \leq -h(0) = d$. So the drift condition becomes $\mathrm{E}_x V(X_1) \leq C + |\theta|d + |\gamma||x|$ proving Proposition 3.

Proof of Proposition 4, Case 2

Using Jensen's inequality and the fact that $\mathrm{P}_x(Y_0 < c) \xrightarrow{x \to \infty} 0$, for all x large enough,

$$\mathrm{E}_x[h(Y_0^*)] \leq h(\mathrm{E}_x[Y_0\mathbf{1}_{Y_0 \geq c}] + c\mathrm{P}_x(Y_0 < c)) = h(\mathrm{E}_x[Y_0] - \mathrm{E}_x[Y_0\mathbf{1}_{Y_0 < c}] + c\mathrm{P}_x(Y_0 < c)).$$

Using a similar argument of Case 1 above, we see that the last two terms in the argument of h converge to 0 as $x \to \infty$. Hence, for (S-3) we have that for any $\varepsilon > 0$ we can find $M > 0$ so that, for all $x > M$, $\mathrm{E}_x[h(Y_0^*)] \leq x + C$; combining this with (S-7), there exists $M > 0$ such that for $x > M$,

$$\mathrm{E}_x V(X_1) \leq C + |\phi|V(x) + |\gamma||x| + |\theta|\mathrm{E}_x|h(Y_0^*) - \bar{g}(\mu)|.$$

It remains to show that the final term in this equation is small relative to the linear (in $V(x)$) term as $x \to \infty$. It is worth noting that the map * does not affect the results for Proposition 5 because

$$\begin{aligned} \mathrm{E}_x(Y_0^*) &= \mathrm{E}_x[Y_0 \mathbf{1}_{Y_0 \geq c}] + \mathrm{E}_x[c \mathbf{1}_{Y_0 < c}] = \mathrm{E}_x(Y_0) - \mathrm{E}_x[Y_0 \mathbf{1}_{Y_0 < c}] + \mathrm{E}_x[c \mathbf{1}_{Y_0 < c}] \\ &= \mu + \mathrm{E}_x[(c - Y_0)\mathbf{1}_{Y_0 < c}] \leq \mu + \mathrm{E}_x[c \mathbf{1}_{Y_0 < c}] \leq \mu + c \end{aligned} \tag{S-8}$$

and $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0^*)] \leq h[\mathrm{E}_x(Y_0^*)] \leq h(\mu + c) \leq h(\mu) + h(c) = h(\mu)$ due to the concavity of h. Hence, the proof follows in almost identical fashion to the proof of this result in Case 1. We omit the details.

Proof of Proposition 2, 3 and 4, Case 3

Assume again $h(c) = 0$. Since $h(Y_0^*) \in [h(c), h(a-c)]$. If $\bar{g}(\mu) = g(\mu)$

$$\begin{aligned} \mathrm{E}_x V(X_1) = \mathrm{E}_x|\alpha + (\phi + \theta)h(Y_0^*) - \theta\tilde{g}(x) + \gamma x| &\leq |\alpha| + |\phi + \theta|\mathrm{E}_x|h(Y_0^*)| + |\gamma - \theta||x| \\ &\leq |\alpha| + |\phi + \theta|h(a-c) + |\gamma - \theta||x|. \end{aligned}$$

Propositions 2-4 follow immediately. If $\bar{g}(\mu) \neq g(\mu)$

$$\begin{aligned} \mathrm{E}_x V(X_1) &\leq |\alpha| + |\phi + \theta|\mathrm{E}_x|h(Y_0^*)| + |\theta||\tilde{g}(x)| + |\gamma||x| \\ &\leq |\alpha| + |\phi + \theta|h(a-c) + |\theta||\tilde{g}(x)| + |\gamma||x|. \end{aligned} \tag{S-9}$$

Proposition 2 follows immediately. We prove Propositions 3 and 4. If $\bar{g}(\mu) = \mathrm{E}_x[h(Y_0^*)]$, equation (S-9) will be $\mathrm{E}_x V(X_1) \leq C + |\theta||\mathrm{E}_x[h(Y_0^*)]| + |\gamma||x| \leq C + |\theta|h(a-c) + |\gamma||x|$. Then, if $\bar{g} \equiv h$, we have that $|\tilde{g}(x)| = |h(\mu)| = h(\mu) < h(a)$, if $h(\mu) > 0$ and $|h(\mu)| = -h(\mu) < -h(0)$ if $h(\mu) < 0$, where $h(a) = \sup_{\mu \in (0,a)} h(\mu)$ and $h(0) = \inf_{\mu \in (0,a)} h(\mu)$. Finally, the drift condition is $\mathrm{E}_x V(X_1) \leq C + |\gamma||x|$ and this completes the proof of Lemma 3. □

Note that Lemma 3 is sufficient to establish stationary conditions for the Case 1, since it involves a continuous-valued process Y_t so the respective chain $X_t = g(\mu_t)$ is φ-irreducible. Resort equation (3.15) from the main paper

$$X_t = \alpha + \gamma\, X_{t-1} + \phi\, h(Y_{t-1}^*) + \theta\left[h(Y_{t-1}^*) - \bar{g}(\mu_{t-1})\right], \tag{S-10}$$

Lemma 4. *The chain $\{X_t\}_{t \in T}$ defined in equation* (S-10) *is weak Feller.*

Proof of Lemma 4

Proof. Define $X_t = g(\mu_t)$ and $X_0 = x$. Let $X_t(x)$ denote the random variable X_t conditional to $X_0 = x$ and $Y_t(x)$ denote the random variable Y_t conditional to $\mu_0 = \mu$. From (S-10) 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 g^{-1} is continuous, $Y_0(g^{-1}(x)) \Rightarrow Y_0(g^{-1}(x'))$ as $x \to x'$. Since the * that maps Y_0 to the domain of h is continuous, it follows that $Y_0^*(g^{-1}(x)) \Rightarrow Y_0^*(g^{-1}(x'))$ as $x \to x'$. Since h is continuous, we have that $h(Y_0^*(g^{-1}(x))) \Rightarrow h(Y_0^*(g^{-1}(x')))$. Since $\tilde{g}(x)$ is continuous, we have that $\tilde{g}(x) \Rightarrow \tilde{g}(x')$. So $X_1(x) \Rightarrow X_1(x')$ as $x \to x'$, showing the weak Feller property. □

For Case 2 and 3, consider the assumption (A3.2):

1. If $\bar{g}(\mu_t) = g(\mu_t)$, $|\gamma - \theta| < 1$
2. If $\bar{g}(\mu_t) \neq g(\mu_t)$ and $|\tilde{g}'(x)| \leq 1$, $|\gamma| + |\theta| < 1$.

Lemma 5. *Assume that Lemma 3, Lemma 4,* (A3.2) *and* (A4) *hold. Then, $\{X_t\}_{t \in T}$ is asymptotic strong Feller.*

Proof of Lemma 5

Proof. When $g \equiv \bar{g}$, it follows from equation (S-10) that $X_1(z) = \alpha + \phi h(Y_0^*(z)) + \theta[h(Y_0^*(z)) - \tilde{g}(z)] + \gamma z$. 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 Roberts and Rosenthal (Prop. 3(g), 2004) 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 to $g(Y_0^*(w)) = g(Y_0^*(z))$ or $h(Y_0^*(w)) = h(Y_0^*(z))$ is equivalent. Therefore, the probability that the chains couple at $t = 1$:

$$\mathrm{P}(g(Y_1^*(w)) = g(Y_1^*(z)) | h(Y_0^*(w)) = h(Y_0^*(z))) > 1 - \left\|\pi_{X_1(z)}(\cdot) - \pi_{X_1(w)}(\cdot)\right\|_{TV} \tag{S-11}$$

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 (S-11): $1 - B|z - w| \sum_{t=0}^{\infty}(|\gamma - \theta|)^t = 1 - \frac{|z-w|B}{1-|\gamma-\theta|}$ where the equality holds by assumption (A3.2). The rest of the proof for the asymptotic strong Feller property follows Matteson et al. (p. 819, 2011). 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 $L \leq 1$, we obtain $|X_1(w) - X_1(z)| \leq (|\theta| + |\gamma|)|z - w|$. Hence, it is immediate to see that the proof for the former case ($\bar{g} \equiv g$) is valid also here by substituting $|\theta| + |\gamma|$ to $|\gamma - \theta|$. This completes the proof. □

Lemma 6. *If* (A3) *hold, then there exists a reachable point* x_0 *for the chain* (S-10).

The condition (A3) is obtained by unifying assumptions (A3.1) and (A3.2).

Proof of Lemma 6

Proof. We show the existence of a reachable point for $\{X_t\}_{t \in T}$ 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 without loss of generality (w.l.o.g.) $h(0) = 0$ (which simply change the value of the constant α). The model (S-10) could be written as

$$x_t = \alpha + \gamma x_{t-1} + (\theta + \phi)h(Y_{t-1}^*) - \theta\tilde{g}(x_{t-1}). \tag{S-12}$$

Let consider the case $Y_t^* = 0$, for $t = 1, \ldots, n$. Hence, by (S-12), $x_t = \alpha + (\gamma - \theta)x_{t-1}$. 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 (A3.2), we have $\lim_{t \to \infty} x_t = x$ so that $\exists\, n$ 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 is immediate to see that $\bar{g}(\mu_t) = 0$, for $t = 1, \ldots, n$ and (S-12) holds as in the previous case, with γ instead of δ, as by (A3.1) follows 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 w.l.o.g. $h(c) = 0$ and (S-12) will be valid as in the former case, with γ instead of δ. □

Proof of Theorem 11

Proof. Theorem 11 follows directly from Lemmata 3-6. More precisely, if (A1)-(A2) and (A3.1) hold, the process $\{X_t\}_{t\in T}$ has at least a stationary distribution. The result is obtained by Lemma 3, Lemma 4 and Theorem 2 in Tweedie (1988). Besides, if (A1)-(A4) hold, the stationary distribution of the process $\{X_t\}_{t\in T}$ is unique. This is immediate by Lemma 5, Lemma 6 and Theorem 3 in Matteson et al. (2011). Finally, by Proposition 8 in Douc et al. (2013), the stationarity of $\{Y_t\}_{t\in T}$ follows directly by the uniqueness of the stationary distribution of $\{X_t\}_{t\in T}$, this completes the proof. □

Proof of equivalence of (A4) and (A5) for Negative Binomial

Proof. For the total variation distance between $d_{TV}(g(Y_t^*(z)), g(Y_t^*(w))) = d_{TV}(Y_t(z), Y_t(w))$, the coupling inequality, as in Thorisson (1995), 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 prove 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

$$\begin{aligned}\mathrm{P}(Y_t(z) \neq Y_t(w)) &= \mathrm{P}(U \neq 0) = 1 - \mathrm{P}(U=0) \\ &\leq 1 - (p_z p_w)^a \sum_{k=0}^{\infty} \binom{a+k-1}{k} [(1-p_z)(1-p_w)]^k \\ &= 1 - \left(\frac{p_z p_w}{1-(1-p_z)(1-p_w)}\right)^a \\ &= 1 - \left(\frac{1}{1+\frac{1-p_z}{p_z}+\frac{1-p_w}{p_w}}\right)^a = 1 - \left(\frac{1}{D}\right)^a \\ &= 1 - \left(\frac{g^{-1}(w) - g^{-1}(z)}{D\left(g^{-1}(w) - g^{-1}(z)\right)}\right)^a \\ &\leq 1 - \left(-\frac{\zeta(z-w)}{D\left(g^{-1}(w) - g^{-1}(z)\right)}\right)^a \qquad \text{(S-13)} \\ &= 1 - \left(\frac{\zeta(z-w)}{D\left(g^{-1}(z) - g^{-1}(w)\right)}\right)^a \\ &\leq 1 - \left(\frac{\zeta(z-w)}{aD^*}\right)^a \qquad \text{(S-14)}\end{aligned}$$

where $D \geq 1$ and and $D\left(g^{-1}(z) - g^{-1}(w)\right) = D_1$. In equation (S-14) we put $D^* = \max\{D, D_1\}$. The inequality (S-13) holds because the function $g^{-1}(\cdot)$ is Lipschitz with constant ζ. Then, (S-14) 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. □

Proof of Corollary 1

Proof. 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 Proposition 2, Case 1 is left unchanged. In Proposition 4 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 proofs still hold with a modification of the constants C (Proposition 5 included). The same thing (with signs inverted) holds for Proposition 3, provided that ν is decreasing w.r.t μ as $x < -M$. For Case 2, Propositions 2 and 4 hold as before. For Proposition 3 we see that $x < -M$ and $0 < \mu < c$, 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

$$\begin{aligned}
\mathrm{E}_x V(X_1) &\leq C + (|\phi| + |\theta|/\nu)a_1 \mathrm{E}_x[Y_0 \mathbf{1}_{Y_0 \geq c}] + |\theta|/\nu|\tilde{g}(x)| + |\gamma||x| \\
&\leq C_2 + C_1/\nu\mu + |\theta|/\nu|\tilde{g}(x)| + |\gamma||x| \\
&\leq C_2 + C_1/\xi c + |\theta|/\xi|\tilde{g}(x)| + |\gamma||x| \\
&\leq C + |\theta|/\xi(l_{22} + l_{23}c^{r/(2+\delta)}) + |\gamma||x|
\end{aligned}$$

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 provide again the same stationarity condition. Then, Lemma 3 holds also for the chain (3.16) in the main paper.

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 4 holds. Also, in order to prove Theorem 11, 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 of Lemma 5. We omit the details. In general, the choice of function $b(\cdot)$ depends on the nature of the process. For example, in the Poisson data case, we can select the VST as $b(Y_0) = \sqrt{Y_0}$. For Negative Binomial data with known number of failure a the VST $b(Y_0^*) = \sqrt{a}\sinh^{-1}(\sqrt{Y_0/a})$ provides the same result. Instead, Dunsmuir and Scott (2015) suggested to set $\nu_t = 1$ (no scaling) for Case 3 since the term $h(Y_{t-1}) - \bar{g}(\mu_{t-1})$ is already bounded. Finally, as here we are in the case where $\bar{g}(\mu_t) = \mathrm{E}[h(Y_t)|\mathcal{F}_{t-1}]$ the existence of a reachable point does not require any modification of the proof for Lemma 6.
Hence, for the Markov chain (3.16) in the main paper, Corollary 1 holds. □

Insight about conditions (H1)-(H2)

In this section, we verify conditions (H1)-(H2) introduced in Section 3.4.1 of the main paper, for particular cases of interest, to show they hold for a large variety of models and are easily verifiable. Of course, the existence of moments of Y_t cannot be proved directly, as its unconditional distribution is unknown, even though they are quite usual assumptions in the context of ML inference. We focus on the other expectations. For convenience in terms of notation, in this paragraph we write $g^\rho\langle Y_{-\infty:t}\rangle = X_t$, even though the process $g^\rho\langle Y_{-\infty:t}\rangle$ in (3.17) in the main paper is not necessarily the same of that in (3.15).

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 usually Lipschitz continuous. Then, we can write $g^{-1}(X_t) - g^{-1}(0) \leq L|X_t|$ and

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

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

$$\mathrm{E}(\log|A'(X_t)|)_+ \quad \leq \mathrm{E}\left(\log^\star(L|X_t|)\right) + \log^\star|g^{-1}(0)| \leq L\,\mathrm{E}|X_t| + b\,. \tag{S-15}$$

So the expectation in (S-15) is finite because the expectation of X_t is finite when $\mathrm{E}|Y_t| < \infty$, see the proof of (B-4) in the Appendix. 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 6, 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 verify 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 $\mathrm{E}\left[A'(X_t)^4\right] = \mathrm{E}\left[A''(X_t)^4\right] \leq \mathrm{E}(Y_t^4) < \infty$.

The verification of conditions (H1)-(H2) for non-canonical link function $g(\cdot)$ clearly depends on its specific shape. We make here some relevant examples. Suppose one wants to model the expectation μ_t linearly as in (3.8) of the main paper, with a Poisson distribution coming from (3.1) of the main paper; 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 $(\alpha, \gamma, \phi, \theta) > 0$, than $X_t = \mu_t \geq \alpha$, completing the proof.

Another common model used in the literature with non-canonical link function is (3.9) for the Negative Binomial (3.10) in the main paper; it is derived by (3.1) in the main paper when $d(Y_t) = \log\frac{\Gamma(\nu+Y_t)}{\Gamma(Y_t+1)\Gamma(\nu)}$, $A(X_t) = -\nu\log\left(\frac{\nu}{\nu+\mu_t}\right) = \nu\log(\nu + e^{X_t}) - \nu\log(\nu)$ and $f(X_t) = \log\left(\frac{\mu_t}{\nu+\mu_t}\right) = X_t - \log(\nu + e^{X_t})$. We know that $\nu > 0$, hence $\mathrm{E}[A'(X_t)^4] = \mathrm{E}[(\frac{\nu e^{X_t}}{\nu+e^{X_t}})^4] \leq \nu^4 < \infty$ and $\mathrm{E}[A''(X_t)^4] = \mathrm{E}[(\frac{\nu^2 e^{X_t}}{(\nu+e^{X_t})^2})^4] \leq \exp(\nu) < \infty$. In the same fashion $f'(X_t)^4 = (\frac{\nu}{\nu+e^{X_t}})^4 \leq 1$ and $f''(X_t)^4 = (\frac{\nu e^{X_t}}{(\nu+e^{X_t})^2})^4 \leq 1$, which posses finite expectations.

Proof of Theorem 14

Proof. The proof of the theorem is based on Douc et al. (Thr. 4.2, 2017), and requires to prove that all the assumptions therein, (A1), (A4), (A5) and (A7), hold when the assumptions of Theorem 14 hold. First of all, note that (A1) is satisfied for the stationarity of Y_t and (A4) is assumed in Theorem 14. 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 (3.19) in the main paper by following the line of Douc et al. (2017) the following assumptions are required to hold true.

(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) $\lim_{t\to\infty}\left\|\chi^{\rho_\star}\left(g^{\bullet}\langle Y_{1:t-1}\rangle(x),Y_t\right)-\chi^{\rho_\star}\left(g^{\bullet}\langle Y_{-\infty:t-1}\rangle,Y_t\right)\right\|=0$,where $\|\cdot\|$ is any norm on $\mathbb{R}^4$.

(ii) $\lim_{t\to\infty}\sup_{\rho\in B(\rho_\star,\epsilon)}\left\|K^{\rho}\left(g^{\bullet}\langle Y_{1:t-1}\rangle(x),Y_t\right)-K^{\rho}\left(g^{\bullet}\langle Y_{-\infty:t-1}\rangle,Y_t\right)\right\|=0$,where $\|\cdot\|$ denote here any norm on 4×4 matrices with real entries.

(iii) $\mathrm{E}\left[\left\|\chi^{\rho_\star}\left(g^{\bullet}\langle Y_{-\infty:0}\rangle,Y_1\right)\right\|^2\right]<\infty\,,\quad \mathrm{E}\left[\sup_{\rho\in B(\rho_\star,\epsilon)}\left\|K^{\rho}\left(g^{\bullet}\langle Y_{-\infty:0}\rangle,Y_1\right)\right\|\right]<\infty.$

Intuitively, (A7) implies that the score function and the information matrix of the data can be approximated by the infinite past of the process. Besides, all these quantities are assumed to exist. 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 $\chi^{\rho}(\cdot,\cdot)=\left[\chi^{\alpha}(\cdot,\cdot),\chi^{\phi}(\cdot,\cdot),\chi^{\gamma}(\cdot,\cdot),\chi^{\theta}(\cdot,\cdot)\right]'$. We compute the derivatives

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

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

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

where we have made implicit $r_j^\star=\gamma_\star-\theta_\star c_j=r_j$ to avoid excesses in the notation. The expressions for the other derivatives are stored in the dedicated section below. An analogous result is found for $\chi^{\rho_\star}\left(g^{\bullet}\langle Y_{-\infty:t-1}\rangle,Y_t\right)$. We show the proof only for one derivative, it is easy to check that the others can be shown in a similar manner. Consider

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

and then

$$\begin{aligned}
&\left|\chi^{\gamma_\star}\left(g^{\bullet}\langle Y_{1:t-1}\rangle(x),Y_t\right)-\chi^{\gamma_\star}\left(g^{\bullet}\langle Y_{-\infty:t-1}\rangle,Y_t\right)\right|\\
&=|Y_t|\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right|\left|f'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right]-f'\left[g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right]\right|+\\
&\quad+|Y_t|\left|f'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right]\right|\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}-\frac{\partial g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle}{\partial\gamma_\star}\right|\\
&\quad+\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right|\left|A'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right]-A'\left[g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right]\right|+ \qquad \text{(S-16)}\\
&\quad+\left|A'\left[g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right]\right|\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}-\frac{\partial g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle}{\partial\gamma_\star}\right|. \qquad \text{(S-17)}
\end{aligned}$$

Now let us verify that

$$\begin{aligned}
\left|\tfrac{\partial g^{\rho_\star}\langle Y_{-\infty:0}\rangle}{\partial\gamma}\right| &\le|\alpha|\sum_{j=0}^{\infty}\tilde{r}^j\sum_{i=0}^{j-1}\frac{1}{r_-}+|\phi+\theta|\sum_{j=0}^{\infty}\tilde{r}^j\left|h(Y^*_{-j})\right|\sum_{i=0}^{j-1}\frac{1}{r_-}\\
&=\tilde{\alpha}\sum_{j=0}^{\infty}\frac{\tilde{r}^j}{r_-}j+\tilde{\delta}\sum_{j=0}^{\infty}\frac{\tilde{r}^j}{r_-}j\left|h(Y^*_{-j})\right|=\frac{\partial\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\gamma}<\infty \qquad \text{(S-18)}
\end{aligned}$$

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

$$\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle}{\partial\gamma}\right| \leq \frac{\partial \hat{g}\langle Y_{1:t-1}\rangle}{\partial\gamma} < \infty. \tag{S-19}$$

Now the difference

$$\begin{aligned}
\left|\tfrac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star} - \tfrac{\partial g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle}{\partial\gamma_\star}\right| &\leq |\alpha_\star| \sum_{l=0}^{\infty} \frac{\tilde{r}^{t+l-1}}{r_-}(t+l-1) + \\
&+ |\phi_\star + \theta_\star| \sum_{l=0}^{\infty} \frac{\tilde{r}^{t+l-1}}{r_-}(t+l-1)\left|h(Y_l^*)\right| + \frac{\tilde{r}^{t-1}}{r_-}(t-1)\left|x\right| \\
&\leq \tilde{r}^{t-1}\left(\tilde{\alpha}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-}l + \tilde{\delta}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-}l\left|h(Y_{-l}^*)\right|\right) + \\
&\tilde{r}^{t-1}(t-1)\left(\tilde{\alpha}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-} + \tilde{\delta}\sum_{l=0}^{\infty}\frac{\tilde{r}^l}{r_-}\left|h(Y_{-l}^*)\right| + \frac{|x|}{r_-}\right) \\
&= \tilde{r}^{t-1}\frac{\partial\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\gamma} + \tilde{r}^{t-1}(t-1)\left(\frac{\hat{g}\langle Y_{-\infty:0}\rangle}{r_-} + \frac{|x|}{r_-}\right) \xrightarrow{t\to\infty} 0
\end{aligned}$$

almost surely, so that (S-17) tends to 0 as $t \to \infty$ according to Douc et al. (2013, Lem. 34), (H1) and equation (S-18). An application of the mean value theorem allows to rewrite equation (S-16) as

$$\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right| |A''(C_{t-1})|\left|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right|,$$

which tends to 0 as $t \to \infty$ for the same reason in (B-6) in the appendix if the following expectation is finite

$$\mathrm{E}\left(\log\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right||A''(C_{t-1})|\right)_+ = \mathrm{E}\left(\log\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right|\right)_+ + \mathrm{E}\left(\log|A''(C_{t-1})|\right)_+ \tag{S-20}$$

The first term of (S-20), $\mathrm{E}\left(\log\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right|\right)_+ \leq \mathrm{E}\left|\frac{\partial g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)}{\partial\gamma_\star}\right| < \infty$ is finite, since, for (H2), the expectation of (S-19) is finite. The proof in the second term of (S-20) 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). Consider

$$\begin{aligned}
\mathrm{E}\left(\log|A''(C_{t-1})|\right)_+ &= \mathrm{E}\left(\log\frac{|A'(g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle) - A'(g^{\rho_\star}\langle Y_{1:t-1}\rangle(x))|}{|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|}\right)_+ \\
&\leq M + \mathrm{E}\left(-\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|\right)_+ \\
&\leq M - \mathrm{E}\left(\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|\right)_- \\
&= M - \frac{1}{2}\mathrm{E}\left(|\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)||\right) + \\
&\quad + \frac{1}{2}\mathrm{E}\left(\log|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle - g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)|\right) \\
&\leq M + \frac{1}{2}\mathrm{E}\left|g^{\rho_\star}\langle Y_{-\infty:t-1}\rangle\right| + \frac{1}{2}\mathrm{E}\left|g^{\rho_\star}\langle Y_{1:t-1}\rangle(x)\right|
\end{aligned} \tag{S-21}$$

which is finite as the expectations of (B-4) and (B-5) in the appendix are for (H1). The same results of (S-16) and (S-17) apply similarly for $f'(\cdot)$, thus are omitted. Hence, (A7)-(i) is proved. We now move to (A7)-(ii). Consider

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

We show the proof only for a single derivative, as the proof of the others is immediate.

$$
\begin{aligned}
&\left|K^{\theta}\left(g^{\bullet}\langle Y_{1:t-1}\rangle(x), Y_t\right) - K^{\theta}\left(g^{\bullet}\langle Y_{-\infty:t-1}\rangle, Y_t\right)\right| \\
&\leq \left[\, |Y_t|\,|f'\left(g^{\rho}\langle Y_{-\infty:t-1}\rangle\right)| + |A'\left(g^{\rho}\langle Y_{-\infty:t-1}\rangle\right)| \,\right] \left|\frac{\partial^2 g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2} - \frac{\partial^2 g^{\rho}\langle Y_{-\infty:t-1}\rangle}{\partial\theta^2}\right| && \text{(S-22)} \\
&\quad + \left|\frac{\partial^2 g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2}\right| \left|A'\left[g^{\rho}\langle Y_{-\infty:t-1}\rangle\right] - A'\left[g^{\rho}\langle Y_{1:t-1}\rangle(x)\right]\right| && \text{(S-23)} \\
&\quad + \left|\frac{\partial^2 g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta^2}\right| |Y_t|\left|f'\left[g^{\rho}\langle Y_{-\infty:t-1}\rangle\right] - f'\left[g^{\rho}\langle Y_{1:t-1}\rangle(x)\right]\right| && \text{(S-24)} \\
&\quad + \left(\frac{\partial g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right)^2 \left|A''\left[g^{\rho}\langle Y_{-\infty:t-1}\rangle\right] - A''\left[g^{\rho}\langle Y_{1:t-1}\rangle(x)\right]\right| && \text{(S-25)} \\
&\quad + \left(\frac{\partial g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right)^2 |Y_t|\left|f''\left[g^{\rho}\langle Y_{-\infty:t-1}\rangle\right] - f''\left[g^{\rho}\langle Y_{1:t-1}\rangle(x)\right]\right| && \text{(S-26)} \\
&\quad + \left[\, |Y_t|\,|f''\left(g^{\rho}\langle Y_{-\infty:t-1}\rangle\right)| + |A''\left(g^{\rho}\langle Y_{-\infty:t-1}\rangle\right)| \,\right] \left|\left(\frac{\partial g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right)^2 - \left(\frac{\partial g^{\rho}\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right)^2\right|.
\end{aligned}
$$

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 that the first element (S-22) tends to 0 as $t\to\infty$ for (H1), by Douc et al. (2013, Lem. 34). The same holds for the elements (S-23) and (S-24) since (S-20) 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 (S-25) and (S-26) also tend to 0 as $t\to\infty$ because of Douc et al. (2013, 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 analogue to (S-21). Finally, it follows that also the last element tends to 0 as $t\to\infty$ for (H1), by Douc et al. (2013, Lem. 34), because it can be rewritten as

$$
\begin{aligned}
\left|\left(\tfrac{\partial g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right)^2 - \left(\tfrac{\partial g^{\rho}\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right)^2\right| \quad &\leq \left|\frac{\partial g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta}\right|\left|\frac{\partial g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta} - \frac{\partial g^{\rho}\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right| \\
&\quad + \left|\frac{\partial g^{\rho}\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right|\left|\frac{\partial g^{\rho}\langle Y_{1:t-1}\rangle(x)}{\partial\theta} - \frac{\partial g^{\rho}\langle Y_{-\infty:t-1}\rangle}{\partial\theta}\right|
\end{aligned}
$$

and this completes the proof for (A7)-(ii). It remains to show (A7)-(iii):

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

where the inequality is obtained by substituting the corresponding equations for the derivatives,

$$
\begin{aligned}
\textstyle\sum_{i=1}^{4}\left(\frac{\partial\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\rho_i}\right)^2 = \quad & \frac{1}{(1-\tilde{r})^2} + 2\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\tilde{r}^{j+i}h(Y_{-j}^*)h(Y_{-i}^*) + 2\frac{\tilde{\alpha}}{r_-^2}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\tilde{r}^{j+i}(ji)h(Y_{-i}^*) + \\
& + 2\frac{\tilde{\delta}}{r_-^2}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\tilde{r}^{j+i}h(Y_{-j}^*)h(Y_{-i}^*) + 4\frac{\tilde{\alpha}\tilde{\delta}}{r_-^2}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\tilde{r}^{j+i}(ji)h(Y_{-i}^*) + \\
& + 2\frac{\tilde{\alpha}}{r_-^2}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\tilde{r}^{j+i}jh(Y_{-i}^*) + 2\frac{\tilde{\delta}}{r_-^2}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\tilde{r}^{j+i}jh(Y_{-j}^*)h(Y_{-i}^*)\,,
\end{aligned}
$$

where, for the Hölder's inequality and (H2),

$$\mathrm{E}\left[Y_1^2 f'\left[g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right]^2 h(Y_{-j}^*)h(Y_{-i}^*)\right] \leq \sqrt{\mathrm{E}\left[Y_1^4\right]}\sqrt{\mathrm{E}\left[f'\left[g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right]^4\right]}\sqrt{\mathrm{E}\left[h(Y_{-j}^*)^2 h(Y_{-i}^*)^2\right]},$$

which is finite. The same is true for $\mathrm{E}\left[A'\left[g^{\rho_\star}\langle Y_{-\infty:0}\rangle\right]^2 h(Y_{-j}^*)h(Y_{-i}^*)\right]$. This proves that the expectation of the score squared is finite by (H2). Analogously, the Hessian

$$\begin{aligned}
\left\|K^\rho\left(g^\bullet\langle Y_{-\infty:0}\rangle, Y_1\right)\right\| &\leq \left(|Y_1|\left|f'\left[g^\rho\langle Y_{-\infty:0}\rangle\right]\right| + \left|A'\left[g^\rho\langle Y_{-\infty:0}\rangle\right]\right|\right)\sum_{j=1}^{4}\sum_{i=1}^{4}\left|\frac{\partial^2\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\rho_j\rho_i}\right| \\
&\leq \left(|Y_1|\left|f''\left[g^\rho\langle Y_{-\infty:0}\rangle\right]\right| + \right. \\
&\qquad \left.\left|A''\left[g^\rho\langle Y_{-\infty:0}\rangle\right]\right|\right)\sum_{j=1}^{4}\sum_{i=1}^{4}\left|\frac{\partial\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\rho_j}\frac{\partial\hat{g}\langle Y_{-\infty:0}\rangle}{\partial\rho_j}\right|
\end{aligned}$$

provides a finite expectation for Hölder's inequality and (H2), completing the proof. □

Simulation studies

Finite sample results

In this section, the numerical results concerning the finite sample properties discussed in Section 3.4.2 are presented. Table S-1 summarises the estimation results for the GLARMA model when the data come from a Bernoulli distribution. Table S-2 and S-3 show the outcome of simulations for GARMA and log-AR models performed on data generated from Geometric distribution in (3.10), but with Poisson distribution fitted instead (QMLE). All the results are based on $s = 1000$ replications, with different configuration of the parameters and increasing sample size $n = (200, 500, 1000, 2000)$. The first row 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 limit 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. In Table S-1 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 Table S-2 and S-3, where the QMLE is fitted. 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.

Table S-1: Simulations for GLARMA(1,1); $Y_t|\mathcal{F}_{t-1} \sim Be(p_t)$, $s = 1000$.

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

Table S-2: Simulations QMLE of Poisson GARMA(1,1); $Y_t|\mathcal{F}_{t-1} \sim Geom(p_t),\ s = 1000$.

n		α	ϕ	θ	α	ϕ	θ	α	ϕ	θ
	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
200	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
	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
500	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
	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
1000	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 S-3: Simulations QMLE of Poisson log-AR(1); $Y_t|\mathcal{F}_{t-1} \sim Geom(p_t)$, $s = 1000$.

n		α	ϕ	γ	α	ϕ	γ
	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
200	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
	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
500	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
	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
1000	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

Model selection

In this section we investigate the model selection on a simulation study. We simulate the first order log-AR, GARMA and GLARMA models, as in Section 3.5.2 of the main paper, 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 $S = 1000$ and sample sizes $n = (250, 500, 1000)$. 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, 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 percent 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 S-4 (results for the BIC are identical).

For the Poisson, 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 works very well also for the Binomial distribution.

Table S-4: 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

Applications

This section includes additional results on the applications discussed in Section 3.5. In particular, we include two plots related to Probability Integral Transform (PIT) in (3.23) of the main paper and the tables on the predictive performance for both the hurricane and Escherichia coli data analysis.

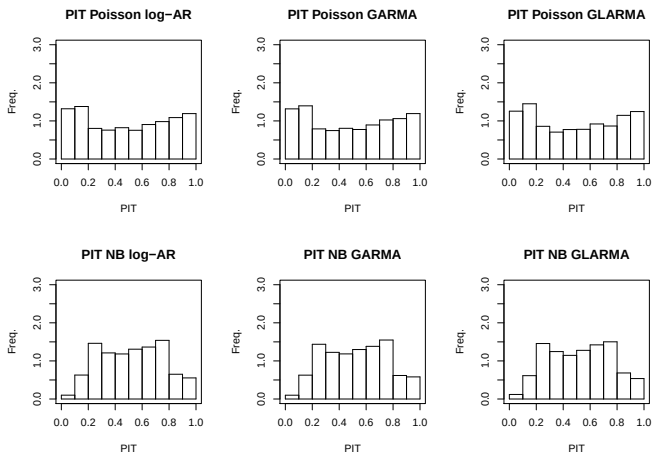

Figure S-1: PIT's for the number of storms. Top: Poisson. Bottom: NB.

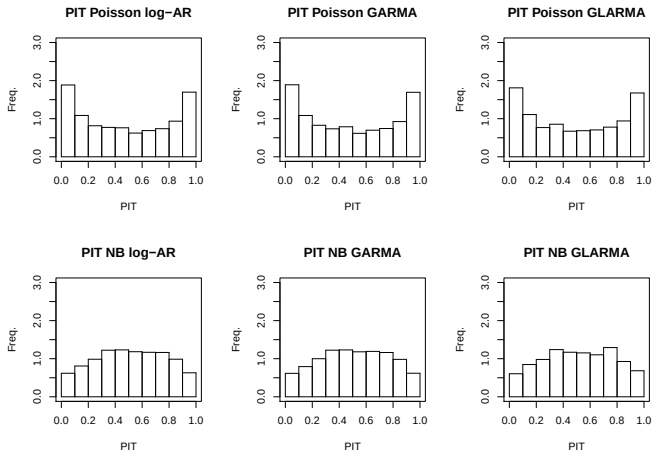

Figure S-2: PIT's for Escheriacoli counts. Top: Poisson. Bottom: NB.

Table S-5: Predictive performance for named storms.

Models	Distribution	logs	qs	sphs	rps
log-AR	Poisson	2.7257	**-0.0775**	**-0.2808**	**2.0320**
	NB	2.8018	-0.0727	-0.2723	2.1235
GARMA	Poisson	2.7293	-0.0774	-0.2807	2.0342
	NB	2.8059	-0.0724	-0.2718	2.1285
GLARMA	Poisson	**2.7247**	-0.0768	-0.2796	2.0384
	NB	2.7927	-0.0735	-0.2736	2.1073

Table S-6: 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	3.5759	-0.0420	-0.2097	3.7347
	NB	3.3286	**-0.0449**	**-0.2127**	**3.6801**

Bibliography

Ahmad, A. and C. Francq (2016). Poisson QMLE of count time series models. *Journal of Time Series Analysis 37*(3), 291–314.

Benjamin, M., R. Rigby, and D. Stasinopoulos (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association 98*(461), 214–223.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological) 26*(2), 211–243.

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

Box, G. E. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Prentice-Hall Inc.

Christou, V. and K. Fokianos (2014). Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis 35*(1), 55–78.

Christou, V. and K. Fokianos (2015). On count time series prediction. *Journal of Statistical Computation and Simulation 85*(2), 357–373.

Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics 8*(2), 93–115.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics 65*(4), 1254–1261.

Davis, R. A., W. T. M. Dunsmuir, and S. B. Streett (2003). Observation-driven models for poisson counts. *Biometrika 90*(4), 777–790.

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

Davis, R. A. and H. Liu (2016). Theory and inference for a class of nonlinear models with application to time series of counts. *Statistica Sinica 26*(4), 1673–1707.

Diaconis, P. and D. Freedman (1999). Iterated random functions. *SIAM 41*(1), 45–76.

Douc, R., P. Doukhan, and E. Moulines (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications 123*(7), 2620 – 2647.

Douc, R., K. Fokianos, and E. Moulines (2017). Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electronic Journal of Statistics 11*(2), 2707–2740.

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

Dunsmuir, W. and D. Scott (2015). The GLARMA package for observation-driven time series regression of counts. *Journal of Statistical Software 67*(7), 1–36.

Englehardt, J. D., N. J. Ashbolt, C. Loewenstine, E. R. Gadzinski, and A. Y. Ayenu-Prah Jr (2012). Methods for assessing long-term mean pathogen count in drinking water and risk management implications. *Journal of Water and Health 10*, 197–208.

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

Fokianos, K., B. Støve, D. Tjøstheim, and P. Doukhan (2020). Multivariate count autoregression. *Bernoulli 26*(1), 471–499.

Fokianos, K. and D. Tjøstheim (2011). Log-linear poisson autoregression. *Journal of Multivariate Analysis 102*(3), 563–578.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B 69*(2), 243–268.

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*.

Li, W. K. (1994). Time series models based on generalized linear models: some further results. *Biometrics 50*(2), 506–511.

Livsey, J., R. Lund, S. Kechagias, and V. Pipiras (2018, 03). Multivariate integer-valued time series with flexible autocovariances and their application to major hurricane counts. *Annals of Applied Statistics 12*(1), 408–431.

Matteson, D. S., D. B. Woodard, and S. G. Henderson (2011). Stationarity of generalized autoregressive moving average models. *Electronic Journal of Statistics 5*, 800–828.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman & Hall.

Meyn, S., R. L. Tweedie, and P. W. Glynn (2009). *Markov Chains and Stochastic Stability* (2 ed.). Cambridge University Press.

Nakagawa, T. and S. Osaki (1975). The discrete weibull distribution. *IEEE Transactions on Reliability 24*, 300–301.

Neumann, M. H. (2011). Absolute regularity and ergodicity of poisson count processes. *Bernoulli 17*(4), 1268–1284.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics 57*, 120–125.

Peluso, A., V. Vinciotti, and K. Yu (2019). Discrete weibull generalized additive model: an application to count fertility data. *Journal of Royal Statistical Society, Series C 68*, 565–583.

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

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

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

Slutsky, E. (1927). The summation of random causes as the source of cyclic processes. *Moscow: Conjuncture Institute 1927*.

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

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

Thorisson, H. (1995). Coupling methods in probability theory. *Scandinavian Journal of Statistics 22*(2), 159–182.

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

Villarini, G., G. A. Vecchi, and J. A. Smith (2010). Modeling the dependence of tropical storm counts in the north atlantic basin on climate indices. *Monthly Weather Review 9*, 353–382.

Walker, G. T. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A 131*(818), 518–532.

Xiao, S., A. Kottas, and B. Sanso (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *Annals of Applied Statistics 9*, 353–382.

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*(636-646), 267–298.

Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics 44*(4), 1019–1031.

Zheng, T., H. Xiao, and R. Chen (2015). Generalized ARMA models with martingale difference errors. *Journal of Econometrics 189*(2), 492 – 506.

Chapter 4

Poisson Network Autoregression

Mirko Armillotta[1,2], Konstantinos Fokianos[2]

[1] *Department of Statistical Sciences, University of Bologna, 41 st. Belle Arti, 40126, Bologna, Italy.*
Email: mirko.armillotta2@unibo.it

[2] *Department of Mathematics and Statistics, University of Cyprus, PO BOX 20537, 1678, Nicosia, Cyprus.*
Email: fokianos@ucy.ac.cy

Abstract

We consider network autoregressive models for count data with a non-random time-varying neighborhood structure. The main methodological contribution is the development of conditions that guarantee stability and valid statistical inference. We consider both cases of fixed and increasing network dimension and we show that quasi-likelihood inference provides consistent and asymptotically normally distributed estimators. The work is complemented by simulation results and a data example.

Keywords: generalized linear models, increasing dimension, link function, multivariate count time series, quasi-likelihood.

4.1 Introduction

The vast availability of integer-valued data, emerging from several real world applications, has motivated the growth of a large literature for modelling and inference about count time series processes. For comprehensive surveys, see Kedem and Fokianos (2002), Davis et al. (2016) and Weiß (2018). Early contributions to the development of count time series models were the Integer Autoregressive models (INAR) Al-Osh and Alzaid (1987); Alzaid and Al-Osh (1990) and observation (Zeger and Liang, 1986) or parameter driven models (Zeger, 1988). The latter classification, due to Cox (1981), will be particular useful as we will be developing theory for Poisson observation-driven models. In this contribution, we appeal to the generalized linear model (GLM) framework, see McCullagh and Nelder (1989), as it provides a natural extension of continuous-valued time series to integer-valued processes. The GLM framework accommodates likelihood inference and supplies a toolbox whereby testing and diagnostics can be also advanced. Some examples of observation-driven models for count time series include the works by Davis et al. (2003), Heinen (2003), Fokianos and Kedem (2004) and Ferland et al. (2006), among others. More recent work includes Fokianos et al. (2009) and Fokianos and Tjøstheim (2011) who develop properties and estimation for a a class of linear and log-linear count time series models. Further related contributions have been appeared over the last years; see Christou

and Fokianos (2014) (for quasi-likelihood inference of negative binomial processes), Ahmad and Francq (2016) (for quasi-likelihood inference based on suitable moment assumptions) and Douc et al. (2013), Davis and Liu (2016), Cui and Zheng (2017) and Douc et al. (2017), among others, for further generalizations of observation-driven models. Theoretical properties of such models have been fully investigated using various techniques; Fokianos et al. (2009) developed initially a perturbation approach, Neumann (2011) employed the notion of β-mixing, Doukhan et al. (2012) (weak dependence approach), Woodard et al. (2011) and Douc et al. (2013) (Markov chain theory without irreducibility assumptions) and Wang et al. (2014) (using e-chains theory; see Meyn and Tweedie (1993)).

Univariate count time series models have been developed and studied in detail, as the previous indicative list of references shows. However, multivariate models, which are necessarily required to be used for network data, are less developed. Studies of multivariate INAR models include those of Latour (1997), Pedeli and Karlis (2011, 2013a,b). Theory and inference for multivariate count time series models is a research topic which is receiving increasing attention. In particular, observation-driven models and their properties are discussed by Heinen and Rengifo (2007), Liu (2012), Andreassen (2013), Ahmad (2016) and Lee et al. (2018). More recently, Fokianos et al. (2020) introduced a multivariate extension of the linear and log-linear Poisson autoregression, as advanced by Fokianos et al. (2009) and Fokianos and Tjøstheim (2011), by employing a copula-based construction for the joint distribution of the counts. These authors employ Poisson processes properties to introduce joint dependence of counts over time. In doing so, they avoid technical difficulties associated with the non-uniqueness of copula for discrete distributions; Genest and Nešlehová (2007). They propose a plausible data generating process which keeps intact, marginally, Poisson processes properties. Further details are given by the review of Fokianos (2021).

The aim of this contribution is to link multivariate observation-driven count time series models with time-varying network data. Such data is increasingly available in many scientific areas (social networks, epidemics, etc.). Measuring the impact of a network structure to a multivariate time series process has attracted considerable attention over the last years; Zhu et al. (2017) for the development of Network Autoregressive models (NAR). These authors have introduced autoregressive models for continuous network data and established associated least squares inference under two asymptotic regimes (a) with increasing time sample size $T \to \infty$ and fixed network dimension N and (b) with both N, T increasing, i.e. $\min\{N, T\} \to \infty$. Significant extension of this work to network quantile autoregressive models has been recently reported by Zhu et al. (2019). Some other extensions of the NAR model include the grouped least squares estimation (Zhu and Pan, 2020) and a network version of the GARCH model, see Zhou et al. (2020) for the case of $T \to \infty$ and fixed network dimension N. Related work was also developed by Knight et al. (2020) who specified a Generalized Network Autoregressive model (GNAR) for continuous random variables, which takes into account different layers of relationships within neighbours of the network. Moreover, the same authors provide an R software for fitting such models. Remark 4 shows that the GNAR model falls in the framework outlined in the present paper.

Following the discussion of Zhu et al. (2017, p. 1116), discrete responses are commonly encountered in real applications and are strongly connected to network data. For example, several data of interest in social network analysis correspond to integer-valued responses. The extension of the NAR model to multivariate count time series is an important theoretical and methodological contribution which is not covered by the existing literature, to the best of our knowledge. The main goal of this work is to fill this gap by specifying linear and log-linear Poisson network autoregressions (PNAR) for count processes and by establishing the two related types of asymptotic inference discussed above. Moreover, the development of all network time series models discussed so far relies strongly on the i.i.d. assumption of the innovations term. Such a condition might not be realistic in many applications. We overcome this limitation by employing the notion of L^p Near epoch dependence (NED), see Andrews (1988), Pötscher and Prucha (1997), and the related concept of α-mixing (Rosenblatt, 1956), (Doukhan, 1994). These notions allow relaxation of the independence assumption as they provide some guarantee of *asymptotic independence* over time. An elaborate and flexible dependence structure among variables, over time and over the nodes composing the network,

is available for all models we consider due to the definition of a full covariance matrix, where the dependence among variables is captured by the copula construction introduced in Fokianos et al. (2020).

For the continuous-valued case, Zhu et al. (2017) employed a simple ordinary least square (OLS) estimation combined with specific properties imposed on the adjacency matrix for the estimation of unknown parameters. However, this method is not applicable to general time series models. In our case, estimation is carried out by using quasi-likelihood methods; see Heyde (1997), for example. When the network dimension N is fixed and the inference with $T \to \infty$ is performed, the standard results already available for Quasi Maximum Likelihood Estimation (QMLE) of Poisson stationary time series, as presented in Fokianos et al. (2009), Fokianos and Tjøstheim (2011) and Fokianos et al. (2020), among others, are also established for the PNAR(p) model. However, the asymptotic properties of the estimators rely on the convergence of sample means to the related expectations due to the ergodicity of a stationary random process $\{\mathbf{Y}_t : t \in \mathbb{Z}\}$ (or a perturbed version of it). The stationarity of an N-dimensional time series, with $N \to \infty$, is still an open problem and it is not clear how it can be achieved. As a consequence, all the results involved by the ergodicity of the time series are unavailable in the increasing dimension case. In the present contribution, this problem is overcome by providing an alternative proof, based on the laws of large numbers for L^p-NED processes of Andrews (1988). Our method requires only the stationarity of the process $\{\mathbf{Y}_t : t \in \mathbb{Z}\}$.

The paper is organized as follows: Section 4.2 discusses the PNAR(p) model specification for the linear and the log-linear case, with lag order p, and the related stability properties. Moreover, a discussion about the empirical structure of the models is provided for the linear first order model ($p = 1$). In Section 4.3, the quasi-likelihood inference is established, showing consistency and asymptotic normality of the quasi maximum likelihood estimator (QMLE) for the two types of asymptotics $T \to \infty$ and $\min\{N, T\} \to \infty$. Section 4.4 discusses the results of a simulation study and an application on real data. The paper concludes with an Appendix containing all the proofs of the main results, the specification of the first two moments for the linear PNAR model, and some further discussion about empirical aspects of the log-linear PNAR(1) model as well as the simulation results.

Notation: We denote $|\mathbf{x}|_r = (\sum_{j=1}^p |x_j|^r)^{1/r}$ the l^r-norm of a p-dimensional vector $\mathbf{x}$. If $r = \infty$, $|\mathbf{x}|_\infty = \max_{1\le j\le p} |x_j|$. Let $\|\mathbf{X}\|_r = (\sum_{j=1}^p \mathrm{E}(|X_j|^r))^{1/r}$ the L^r-norm for a random vector $\mathbf{X}$. For a $q \times p$ matrix $\mathbf{A} = (a_{ij})$, $i = 1, \dots, q, j = 1, \dots, p$, denotes the generalized matrix norm $|||\mathbf{A}|||_r = \max_{|\mathbf{x}|_r=1} |\mathbf{A}\mathbf{x}|_r$. If $r = 1$, then $|||\mathbf{A}|||_1 = \max_{1\le j\le p} \sum_{i=1}^q |a_{ij}|$. If $r = 2$, $|||\mathbf{A}|||_2 = \rho^{1/2}(\mathbf{A}^T\mathbf{A})$, where $\rho(\cdot)$ is the spectral radius. If $r = \infty$, $|||\mathbf{A}|||_\infty = \max_{1\le i\le q} \sum_{j=1}^p |a_{ij}|$. If $q = p$, then these norms are matrix norms.

4.2 Models

We consider a network with N nodes (network size) and index $i = 1, \dots N$. The structure of the network is completely described by the adjacency matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N\times N}$ where $a_{ij} = 1$ if there is a directed edge from i to j, $i \to j$ (e.g. user i follows j on Twitter), and $a_{ij} = 0$ otherwise. However, undirected graphs are allowed ($i \leftrightarrow j$). The structure of the network is assumed nonrandom. Self-relationships are not allowed $a_{ii} = 0$ for any $i = 1, \dots, N$, this is a typical assumption, and it is reasonable for various real situations, e.g. social media. For details about the definition of social networks see Wasserman et al. (1994), Kolaczyk and Csárdi (2014). Let us define a certain count variable $Y_{i,t} \in \mathbb{R}$ for the node i at time t. We want to assess the effect of the network structure on the count variable $\{Y_{i,t}\}$ for $i = 1, \dots, N$ over time $t = 1, \dots, T$.

In this section, we study the properties of linear and log-linear models. We initiate this study by considering a simple, yet illuminating, case of a linear model of order one and then we consider the more general case of p'th order model. Finally, we discuss log-linear models. In what follows, we denote by $\{\mathbf{Y}_t = (Y_{it},\, i = 1, 2 \dots N,\, t = 0, 1, 2 \dots, T)\}$ an N-dimensional vector of count time series with $\{\boldsymbol{\lambda}_t = (\lambda_{it},\, i = 1, 2 \dots N,\, t = 1, 2 \dots, T)\}$ be the corresponding N-dimensional intensity process vector. Define by $\mathcal{F}_t = \sigma(\mathbf{Y}_s : s \le t)$. Based on the specification of the model, we

assume that $\boldsymbol{\lambda}_t = \mathrm{E}(\mathbf{Y}_t|\mathcal{F}_{t-1})$.

4.2.1 Linear PNAR(1) model

A linear count network model of order 1, is given by

$$Y_{it}|\mathcal{F}_{t-1} \sim Poisson(\lambda_{it}), \quad \lambda_{i,t} = \beta_0 + \beta_1 n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{jt-1} + \beta_2 Y_{it-1}\,, \tag{4.1}$$

where $n_i = \sum_{j \neq i} a_{ij}$ is the out-degree, i.e the total number of nodes which i has an edge with. From the left hand side equation of (4.1), we observe that the process Y_{it} is assumed to be marginally Poisson. We call (4.1) linear Poisson network autoregression of order 1, abbreviated by PNAR(1).

The development of a multivariate count time series model would lead to the specification of a joint distribution, so that the standard likelihood inference and testing procedures can be performed accordingly. Although several alternatives have been proposed in the literature, see the review in Fokianos (2021, Sec. 2), the choice of a suitable multivariate version of the Poisson probability mass function (p.m.f) is far from obvious. In fact, a multivariate Poisson-type p.m.f has a complicated closed form and the associated likelihood inference is theoretically cumbersome and numerically challenging. Furthermore, in many cases, the available multivariate Poisson-type p.m.f. implicitly implies restricted models, which are of limited use in applications (e.g. covariances always positive, constant pairwise correlations). For these reasons, in the present paper the joint distribution of the vector $\{\mathbf{Y}_t\}$ is constructed by following the approach of Fokianos et al. (2020, p. 474), imposing a copula structure on waiting times of a Poisson process. More precisely,

1. Let $\mathbf{U}_l = (U_{1,l}, \ldots, U_{N,l})$, for $l = 1, \ldots, K$ a sample from a N-dimensional copula $C(u_1, \ldots, u_N)$, where $U_{i,l}$ follows a Uniform(0,1) distribution, for $i = 1, \ldots, N$.
2. The transformation $X_{i,l} = -\log U_{i,l}/\lambda_{i,0}$ is exponential with parameter $\lambda_{i,0}$, for $i = 1, \ldots, N$.
3. The process $Y_{i,0} = \max_{1 \leq k \leq K} \left\{ \sum_{l=1}^{k} X_{i,l} \leq 1 \right\}$ is Poisson with parameter λ_0, for $i = 1, \ldots, N$. So, $\mathbf{Y}_0 = (Y_{1,0}, \ldots, Y_{N,0})$ is a set of marginal Poisson processes with mean λ_0.
4. By using the model (4.1), λ_1 is obtained.
5. Return back to step 1 to obtain $\mathbf{Y}_1$, and so on.

The described data generating process ensures all the marginal distributions of the variables Y_{it} to be univariate Poisson, as described in (4.1), while an arbitrary dependence among them is introduced in a flexible and general way. For a comprehensive discussion on the choice of a multivariate count distribution and the comparison between the alternatives proposed, the interested reader can refer to Fokianos (2021).

Model (4.1) postulates that, for every single node i, the marginal conditional mean of the process is regressed on the past count of the variable itself for i and the average count of the other nodes $j \neq i$ which have a connection with i. This model assumes that only the nodes which are directly followed by the focal node i possibly have an impact on the mean process of counts. It is a reasonable assumption in many applications; for example, in a social network the activity of node k, which satisfies $a_{ik} = 0$, does not affect node i. The parameter β_1 is called network effect, as it measures the average impact of node i's connections $n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{jt-1}$. The coefficient β_2 is called momentum effect because it provides a weight for the impact of past count Y_{it-1}. This interpretation is in line with the Gaussian network vector autoregression (NAR) introduced by Zhu et al. (2017) for continuous variables.
For simplicity, we rewrite model (4.1) in a vector form, as in Fokianos et al. (2020),

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \boldsymbol{\beta}_0 + \mathbf{G}\mathbf{Y}_{t-1}\,, \tag{4.2}$$

where $\{\mathbf{N}_t\}$ is a sequence of independent N-variate copula-Poisson process, which counts the number of events in $[0, \lambda_{1,t}] \times \cdots \times [0, \lambda_{1,t}]$. We also define $\boldsymbol{\beta}_0 = \beta_0 \mathbf{1}_N \in \mathbb{R}^N$ with $\mathbf{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^N$ and the matrix $\mathbf{G} = \beta_1 \mathbf{W} + \beta_2 \mathbf{I}_N$ where $\mathbf{W} = \text{diag}\left\{n_1^{-1}, \ldots, n_N^{-1}\right\} \mathbf{A}$ is the row-normalized adjacency matrix, $\mathbf{A} = (a_{ij})$, so $w_i = (a_{ij}/n_i,\, j = 1, \ldots, N)^T \in \mathbb{R}^N$ is the i-th row vector of the matrix $\mathbf{W}$, and $\mathbf{I}_N$ is the identity matrix $N \times N$. Note that the matrix $\mathbf{W}$ is a (row) stochastic matrix, as $|||\mathbf{W}|||_\infty = 1$ (Seber, 2008, Def. 9.16).

To gain intuition for model (4.1), we simulate a network from the stochastic block model (Wang and Wong, 1987); see Figure 4.1. Moments of the linear model (4.1) exist and have a closed form expression; see (C-2). The mean vector of the process has elements $\mathrm{E}(Y_{it})$ which vary between 0.333 to 0.40, for $i = 1, \ldots, N$ whereas the diagonal elements of $\mathrm{Var}(\mathbf{Y}_t)$ take values between 0.364 and 0.678. We take this simulated model as a baseline for comparisons and its correlation structure is shown in the upper-left plot of Figure 4.1. The top-right panel displays the same information but for the case of increasing activity in the network. The bottom panel of the same figure shows the same information as the upper panel but with a more sparse network, i.e. $K = 10$. Increasing the number of relationships among nodes of the network boosts the correlation among the count processes. A more sparse structure of the network does not appear to alter the correlation properties of the process though.

Figure 4.2 shows a substantial increase in the correlation values which is due to the choice of the copula parameter. Interestingly, the intense activity of the network increases the correlation values of the count process. This aspect may be expected in real applications. For the Clayton copula (see lower plots of the same figure) we observe the same phenomenon but the values of the correlation matrix are lower when compare to those of the Gaussian copula. We did not observe any substantial changes for the marginal mean and variances.

Figure 4.3 shows the impact of increasing network and momentum effects. We observe that the network effect is prevalent, as it can be seen from the top-right panel which also shows the block network structure. Significant inflation for the correlation can be also noticed when increasing the momentum effect (bottom-left panel). When increasing the network effect the marginal means vary between 0.333 to 1 and have large variability within the nodes; this is a direct consequence of the block network structure. When increasing the momentum effect, the marginal means take values from 0.5 to 0.667. When both effects grow, the mean values increase and are between 0.5 and 2.

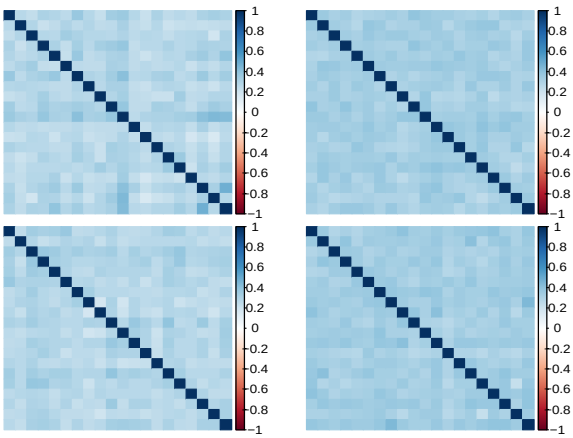

Figure 4.1: Correlation matrix of model (4.1). Top-left: Data are generated by employing a stochastic block model with $K = 5$ and an adjacency matrix $\mathbf{A}$ with elements generated by $\mathrm{P}(a_{ij} = 1) = 0.3N^{-0.3}$, if i and j belong to the same block, and $\mathrm{P}(a_{ij} = 1) = 0.3N^{-1}$, otherwise. In addition, we employ a Gaussian copula with parameter $\rho = 0.5$, $(\beta_0, \beta_1, \beta_2) = (0.2, 0.1, 0.4)^T$, $T = 2000$ and $N = 20$. Top-right plot: Data are generated by employing a stochastic block model with $K = 5$ and an adjacency matrix $\mathbf{A}$ with elements generated by $\mathrm{P}(a_{ij} = 1) = 0.7N^{-0.0003}$ if i and j belong to the same block, and $\mathrm{P}(a_{ij} = 1) = 0.6N^{-0.3}$ otherwise. Same values for β's, T, N and choice of copula. Bottom-left: The same graph, as in the upper-left side but with $K = 10$. Bottom-right: The same graph, as in upper-right side but with $K = 10$.

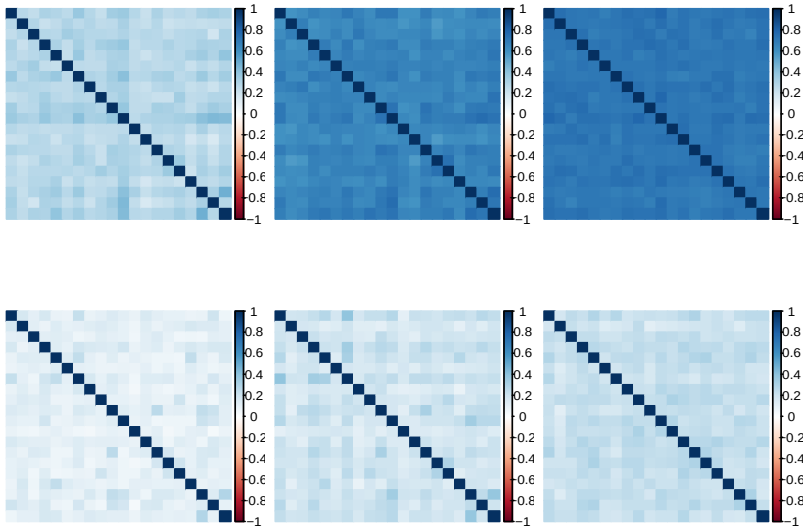

Figure 4.2: Correlation matrix of model (4.1). Top: Data have been generated as in top-left of Figure 4.1 (left), with copula correlation parameter $\rho = 0.9$ (middle) and as in the top-right of Figure 4.1 but with copula parameter $\rho = 0.9$ (right). Bottom: same information as the top plot but data are generated by using a Clayton copula.

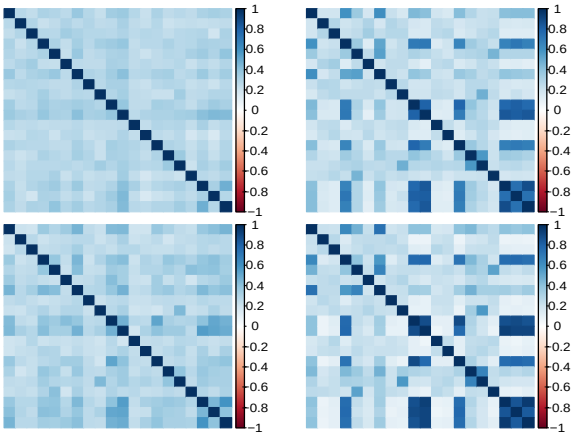

Figure 4.3: Correlation matrix of model (4.1). Data have been generated as in top-left of Figure 4.1 (top-left), higher network effect $\beta_1 = 0.4$ (top-right), higher momentum effect $\beta_2 = 0.6$ (lower-left) and higher network and momentum effect $\beta_1 = 0.3$, $\beta_2 = 0.6$ (lower-right).

4.2.2 Linear PNAR(p) model

More generally, we introduce and study an extension of model (4.1) by allowing Y_{it} to depend on the last p lagged values. We call this the linear Poisson NAR(p) model and its defined analogously to (4.1) but with

$$\lambda_{i,t} = \beta_0 + \sum_{h=1}^{p} \beta_{1h} \left(n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{jt-h} \right) + \sum_{h=1}^{p} \beta_{2h} Y_{it-h} \,, \tag{4.3}$$

where $\beta_0, \beta_{1h}, \beta_{2h} \geq 0$ for all $h = 1 \ldots, p$. If $p = 1$, $\beta_{11} = \beta_1$, $\beta_{22} = \beta_2$ to obtain (4.1). The joint distribution of the vector $\mathbf{Y}_t$ is defined by means of the copula construction discussed in Sec. 4.2.1. Without loss of generality, we can set coefficients equal to zero if the parameter order is different in both terms of (4.3). Its is easy to see that (4.3) can be rewritten as

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \boldsymbol{\beta}_0 + \sum_{h=1}^{p} \mathbf{G}_h \mathbf{Y}_{t-h} \,, \tag{4.4}$$

where $\mathbf{G}_h = \beta_{1h}\mathbf{W} + \beta_{2h}\mathbf{I}_N$ for $h = 1, \ldots, p$ by recalling that $\mathbf{W} = \operatorname{diag}\left\{n_1^{-1}, \ldots, n_N^{-1}\right\} \mathbf{A}$. We have the following result which gives verifiable conditions equivalent to the conditions of Zhu et al. (2017, Thm.1) for continuous values network autoregression.

Proposition 6. Consider model (4.3) (or equivalently (4.4)). Suppose that $\sum_{h=1}^{p}(\beta_{1h} + \beta_{2h}) < 1$. Then the process $\{\mathbf{Y}_t,\ t \in \mathbb{Z}\}$ is stationary and ergodic with $\mathrm{E}\,|\mathbf{Y}_t|_1^r < \infty$ for any $r > 1$ and fixed N.

Proof. The result follows from Debaly and Truquet (2019, Thm. 4), provided that $\rho(\sum_{h=1}^{p} \mathbf{G}_h) < 1$. But $\rho(\sum_{h=1}^{p} \mathbf{G}_h) \leq |||\sum_{h=1}^{p} \mathbf{G}_h|||_\infty \leq \sum_{h=1}^{p} |||\mathbf{G}_h|||_\infty \leq \sum_{h=1}^{p}(\beta_{1h}|||\mathbf{W}|||_\infty + \beta_{2h}) = \sum_{h=1}^{p}(\beta_{1h} + \beta_{2h})$, since $|||\mathbf{W}|||_\infty = 1$ by construction. Therefore we conclude that $\{\mathbf{Y}_t,\ t \in \mathbb{Z}\}$ is a stationary and ergodic process with $\mathrm{E}\,|\mathbf{Y}_t|_1^r < \infty$ for any $r > 1$. □

Some further results about the first and second order properties of model (4.3) are given in the Appendix. Similar results have been recently reported by Fokianos et al. (2020) when there is a feedback in the model. Following these authors, we obtain the same results of Proposition 6 but under stronger conditions. For example, when $p = 1$, we will need to assume either $|||\mathbf{G}|||_1$ or $|||\mathbf{G}|||_2 < 1$ to obtain identical results. The condition $\sum_{h=1}^{p}(\beta_{1h} + \beta_{2h}) < 1$ is more natural and complements the existing work on continuous valued models Zhu et al. (2017). In addition, note

that the copula construction is not used in the proof of Prop. 6 (see also Prop. 8 for log-linear model). However, it is used in Section 4.4.1 where we report a simulation study. It is interesting though this setup is similar to multivariate ARMA models, where the stability conditions are independent of the correlations in the innovation.

Proposition 6 states that all the moments exist finite, for fixed N. A similar results is also proved in Fokianos et al. (2020, Prop. 3.2). The following results state that even when N is increasing all the moments exist and are uniformly bounded. For clarity in the notation, we present the result for the PNAR(1) model, but it can be easily extended to hold true for $p > 1$.

Proposition 7. *Consider the model* (4.1) *and the stationarity condition* $\beta_1 + \beta_2 < 1$. *Then,* $\max_{i\geq 1} \mathrm{E}\,|Y_{it}|^r < C_r < \infty$, *for any* $r \in \mathbb{N}$.

Proof. By (C-2), recall that $\mathrm{E}(Y_{it}) = \mu = \beta_0/(1-\beta_1-\beta_2)$ for all $1 \leq i \leq N$. Then, $\max_{1\leq i\leq N} \mathrm{E}(Y_{it}) = \mu$ and $\lim_{N\to\infty} \max_{1\leq i\leq N} \mathrm{E}(Y_{it}) = \max_{i\geq 1} \mathrm{E}(Y_{it}) \leq \mu = C_1$, using properties of monotone bounded functions. Moreover, $\mathrm{E}(Y_{it}^r|\mathcal{F}_{t-1}) = \sum_{k=1}^r \left\{ {r \atop k} \right\} \lambda_{it}^r$, employing Poisson properties, where $\left\{ {r \atop k} \right\}$ are the Stirling numbers of the second kind. Set $r = 2$. For the law of iterated expectations (Billingsley, 1995, Thm. 34.4), we have that

$$
\begin{aligned}
\max_{1\leq i\leq N} \|Y_{it}\|_2 = \max_{1\leq i\leq N} \left[\mathrm{E}\left(\lambda_{it}^2 + \lambda_{it}\right)\right]^{1/2} &\leq \max_{1\leq i\leq N} \left[\mathrm{E}\left(\beta_0 + \beta_1 \sum_{j=1}^N w_{ij} Y_{jt-1} + \beta_2 \|Y_{it-1}\|_2\right)^2 + \mu\right]^{1/2} \\
&\leq \beta_0 + \beta_1 \max_{1\leq i\leq N} \left(\sum_{j=1}^N w_{ij} \|Y_{jt-1}\|_2\right) + \beta_2 \max_{1\leq i\leq N} \|Y_{it-1}\|_2 + \mu^{1/2} \\
&\leq \beta_0 + (\beta_1+\beta_2) \max_{1\leq i\leq N} \|Y_{it-1}\|_2 + \mu^{1/2} \\
&\leq \frac{\beta_0 + \mu^{1/2}}{1-\beta_1-\beta_2} = C_2^{1/2} < \infty\,,
\end{aligned}
$$

where the last inequality works for the stationarity of the process $\{\mathbf{Y}_t, t \in \mathbb{Z}\}$ and the finiteness of its moments, with fixed N. As $\max_{1\leq i\leq N} \mathrm{E}\,|Y_{it}|^2$ is bounded by C_2, for the same reason above $\max_{i\geq 1} \mathrm{E}\,|Y_{it}|^2 \leq C_2$. Since $\mathrm{E}(Y_{it}^3|\mathcal{F}_{t-1}) = \lambda_{it}^3 + 3\lambda_{it}^2 + \lambda_{it}$, similarly as above

$$
\begin{aligned}
\max_{1\leq i\leq N} \|Y_{it}\|_3 &\leq \beta_0 + (\beta_1+\beta_2) \max_{1\leq i\leq N} \|Y_{it-1}\|_3 + (3\mathrm{E}(\lambda_{it}^2))^{1/3} + \mu^{1/3} \\
&\leq \beta_0 + (\beta_1+\beta_2) \max_{1\leq i\leq N} \|Y_{it-1}\|_3 + (3C_2)^{1/3} + \mu^{1/3} \\
&\leq \frac{\beta_0 + (3C_2)^{1/3} + \mu^{1/3}}{1-\beta_1-\beta_2} = C_3^{1/3} < \infty\,,
\end{aligned}
$$

where the second inequality holds for the conditional Jensen's inequality, and so on, for $r > 3$, the proof works analogously by induction, therefore is omitted. □

4.2.3 Log-linear PNAR models

Recall model (4.1). The network effect β_1 of model (4.1) is typically expected to be positive, see Chen et al. (2013), and the impact of Y_{it-1} is positive, as well. Hence, positive constraints on the parameters are theoretically justifiable as well as practically sound. However, in order to allow a better link to the GLM theory, McCullagh and Nelder (1989), and adding the possibility to insert covariates as well as coefficients which take values on the entire real line and cannot be estimated by a linear model, we propose the following log-linear model, see Fokianos and Tjøstheim (2011):

$$
Y_{it}|\mathcal{F}_{t-1} \sim Poisson(\nu_{i,t}), \quad \nu_{it} = \beta_0 + \beta_1 n_i^{-1} \sum_{j=1}^N a_{ij} \log(1 + Y_{jt-1}) + \beta_2 \log(1 + Y_{it-1})\,, \tag{4.5}
$$

where $\nu_{it} = \log(\lambda_{it})$ for every $i = 1, \ldots, N$. No constraints are required in model (4.5) since $\nu_{it} \in \mathbb{R}$. The interpretation of parameters and additive components remains unchanged. Again, the model can be rewritten in vectorial form, as in the case of model (4.1)

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\nu}_t), \quad \boldsymbol{\nu}_t = \boldsymbol{\beta}_0 + \mathbf{G} \log(\mathbf{1}_N + \mathbf{Y}_{t-1}), \tag{4.6}$$

with $\boldsymbol{\nu}_t \equiv \log(\boldsymbol{\lambda}_t)$, componentwise. Furthermore, we can have a useful approximation by

$$\log(\mathbf{1}_N + \mathbf{Y}_t) = \boldsymbol{\beta}_0 + \mathbf{G} \log(\mathbf{1}_N + \mathbf{Y}_{t-1}) + \boldsymbol{\psi}_t,$$

where $\boldsymbol{\psi}_t = \log(\mathbf{1}_N + \mathbf{Y}_t) - \boldsymbol{\nu}_t$. By lemma A.1 in Fokianos and Tjøstheim (2011) $\mathrm{E}(\boldsymbol{\psi}_t | \mathcal{F}_{t-1}) \to 0$ as $\boldsymbol{\nu}_t \to \infty$, so $\boldsymbol{\psi}_t$ is "approximately" martingale difference sequence (MDS). Moreover, one can define here the martingale difference sequence $\boldsymbol{\xi}_t = \mathbf{Y}_t - \exp(\boldsymbol{\nu}_t)$. We discuss empirical properties of the model (4.5) in the Appendix. More generally, we define the log-linear PNAR(p) by

$$\nu_{it} = \beta_0 + \sum_{h=1}^{p} \beta_{1h} \left(n_i^{-1} \sum_{j=1}^{N} a_{ij} \log(1 + Y_{jt-h}) \right) + \sum_{h=1}^{p} \beta_{2h} \log(1 + Y_{it-h}), \tag{4.7}$$

using the same notation as before. The interpretation of this model is developed along the lines of the linear model. Furthermore,

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\nu}_t), \quad \boldsymbol{\nu}_t = \boldsymbol{\beta}_0 + \sum_{h=0}^{p} \mathbf{G}_h \log(\mathbf{1}_N + \mathbf{Y}_{t-h}), \tag{4.8}$$

where $\mathbf{G}_h = \beta_{1h}\mathbf{W} + \beta_{2h}\mathbf{I}_N$ for $h = 1, \ldots, p$.

Proposition 8. Consider model (4.7) (or equivalently (4.8)). Suppose that $\sum_{h=1}^{p}(|\beta_{1h}| + |\beta_{2h}|) < 1$. Then the process $\{\mathbf{Y}_t,\ t \in \mathbb{Z}\}$ is stationary and ergodic with $\mathrm{E}\,|\mathbf{Y}_t|_1 < \infty$ and there exists $\delta > 0$ such that $\mathrm{E}[\exp(\delta\,|\mathbf{Y}_t|_1^r)] < \infty$ and $\mathrm{E}[\exp(\delta\,|\boldsymbol{\nu}_t|_1^r)] < \infty$ for fixed N.

Proof. The result follows from Debaly and Truquet (2019, Thm. 5), provided that $\left\|\left|\sum_{h=1}^{p} |\mathbf{G}_h|_e\right|\right\|_\infty < 1$, where $|\cdot|_e$ is the elementwise absolute value. But $\left\|\left||\mathbf{G}_h|_e\right|\right\|_\infty \leq |\beta_{1h}| \left\|\left|\mathbf{W}\right|\right\|_\infty + |\beta_{2h}| = |\beta_{1h}| + |\beta_{2h}|$. Therefore we conclude that $\{\mathbf{Y}_t,\ t \in \mathbb{Z}\}$ is a stationary and ergodic process with $\mathrm{E}\,|\mathbf{Y}_t|_1 < \infty$ and there exists $\delta > 0$ such that $\mathrm{E}[\exp(\delta\,|\mathbf{Y}_t|_1^r)] < \infty$ and $\mathrm{E}[\exp(\delta\,|\boldsymbol{\nu}_t|_1^r)] < \infty$. □

Remark 3. Taking into account known time-varying network structures, i.e. $\mathbf{A}_t$, $t = 1, \ldots, T$ denote dynamic adjacency matrices, is of potential interest in applications. In this case, model (4.2) is written as

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \boldsymbol{\beta}_0 + \mathbf{G}_t \mathbf{Y}_{t-1},$$

where $\mathbf{G}_t = \beta_1 \mathbf{W}_t + \beta_2 \mathbf{I}_N$ and $\mathbf{W}_t = \mathrm{diag}\left\{n_{1,t}^{-1}, \ldots, n_{N,t}^{-1}\right\} \mathbf{A}_t$. It is worth noting that $\left\|\left|\mathbf{W}_t\right|\right\|_\infty = 1$, is still true for every $t = 1, \ldots, T$, so $\left\|\left|\mathbf{W}_t\right|\right\|_\infty = \left\|\left|\mathbf{W}\right|\right\|_\infty$, which is the only property required for this matrix, throughout the paper. Even though $\rho(\mathbf{G}_t) < 1$, for every t, Propositions 6 and 8 do not apply. Provided that the model is stationary, all methods and results developed in the present contribution extend straightforwardly to time-varying network structures. To avoid excessive notation, the results reported in the paper are under the condition $\mathbf{W}_t = \mathbf{W}$.

Remark 4. Another suitable extension encompassed in this paper is the GNAR(p) version introduced in Knight et al. (2020, eq. 1) in the context of continuous-valued random variables. This model adds an average neighbour impact for several stages of connections between the nodes of a given network. Define $\mathcal{N}(\{i\}) = \{j \in \{1, \ldots, N\} : i \to j\}$ the set of neighbours of the node i. Then, $\mathcal{N}^{(r)}(i) = \mathcal{N}\left\{\mathcal{N}^{(r-1)}(i)\right\} / \left[\left\{\cup_{q=1}^{r-1} \mathcal{N}^{(q)}(i)\right\} \cup \{i\}\right]$, for $r = 2, 3, \ldots$ is the set of r-stage neighbours of i and $\mathcal{N}^{(1)}(i) = \mathcal{N}(\{i\})$. (So, for example, $\mathcal{N}^{(2)}(i)$ describes the neighbours

of the neighbours of the node i, and so on.) In this case, the row-normalized adjacency matrix have elements $\left(\mathbf{W}^{(r)}\right)_{i,j} = w_{i,j} \times I(j \in \mathcal{N}^{(r)}(i))$, where $w_{i,j} = 1/\text{card}(\mathcal{N}^{(r)}(i))$, $\text{card}(\cdot)$ denotes the cardinality of a set and $I(\cdot)$ is the indicator function. Several C types of edges are allowed in the network. Moreover, time-varying networks can be considered as well. Under the framework, the Poisson GNAR(p) has the following formulation.

$$\lambda_{i,t} = \beta_0 + \sum_{h=1}^{p} \left(\sum_{c=1}^{C} \sum_{r=1}^{s_h} \beta_{1,h,r,c} \sum_{j \in \mathcal{N}_t^{(r)}(i)} w_{i,j,c}^{(t)} Y_{jt-h} + \beta_{2,h} Y_{it-h} \right), \tag{4.9}$$

where s_h is the maximum stage of neighbour dependence for the time lag h. Model (4.9) can be included in the formulation (4.4) by setting $\mathbf{G}_h = \sum_{c=1}^{C} \sum_{r=1}^{s_h} \beta_{1,h,r,c} \mathbf{W}^{(r,c)} + \beta_{2,h} \mathbf{I}_N$. Since it holds that $\sum_{j \in \mathcal{N}^{(r)}(i)} \sum_{c=1}^{C} w_{i,j,c} = 1$, we have $\left\| \sum_{c=1}^{C} \mathbf{W}^{(r,c)} \right\|_\infty = 1$. The time-varying network extension is straightforward, by taking into account Remark 3. Then, all the results of the present contribution apply directly to (4.9). Analogous arguments hold true for the log-linear model (4.7).

4.3 Estimation

4.3.1 Quasi-likelihood inference for fixed N

We approach the estimation problem by using the theory of estimating functions; see Basawa and Prakasa Rao (1980), Zeger and Liang (1986) and Heyde (1997), among others. Let the vector of unknown parameters $\boldsymbol{\theta} = (\beta_0, \beta_{11}, \ldots, \beta_{1p}, \beta_{21}, \ldots, \beta_{2p})^T \in \mathbb{R}^m$, where $m = 2p + 1$. Define the conditional quasi-log-likelihood function for $\boldsymbol{\theta}$:

$$l_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{i=1}^{N} y_{i,t} \log \lambda_{i,t}(\boldsymbol{\theta}) - \lambda_{i,t}(\boldsymbol{\theta}), \tag{4.10}$$

which is the log-likelihood one would obtain if time series modelled in (4.2), or (4.6), would be contemporaneously independent. This simplifies computations but guarantees consistency and asymptotic normality of the resulting estimator. Although the joint copula structure $C(\ldots, \rho)$ and the set of parameters ρ, usually describing its functional form, are not included in the maximization of the "working" log-likelihood (4.10), this does not mean that the inference is carried out under the assumption of independence along the observed process, conditionally on the past $\mathcal{F}_{t-1}$; it can easily be detected from the shape of the conditional information matrix (4.14) below, which takes into account the true conditional covariance matrix of the process $\mathbf{Y}_t$.

Douc et al. (2017), among others, established inference theory for Quasi Maximum Likelihood Estimation (QMLE) for observation driven models. Assuming that there exist a true vector of parameter, say $\boldsymbol{\theta}_0$, such that the mean model specification (4.2) (or equivalently (4.6)) is correct, regardless the true data generating process, then we obtain a consistent and asymptotically normal estimator by maximizing the quasi-log-likelihood (4.10). Denote by $\hat{\boldsymbol{\theta}} \coloneqq \arg\max_{\boldsymbol{\theta}} l_{NT}(\boldsymbol{\theta})$, the QMLE for $\boldsymbol{\theta}$. The score function for the linear model is given by

$$\begin{aligned} \mathbf{S}_{NT}(\boldsymbol{\theta}) &= \sum_{t=1}^{T} \sum_{i=1}^{N} \left(\frac{y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \sum_{i=1}^{T} \frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta}) \Big) = \sum_{t=1}^{T} \mathbf{s}_{Nt}(\boldsymbol{\theta}), \end{aligned} \tag{4.11}$$

where

$$\frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = (\mathbf{1}_N, \mathbf{W}\mathbf{Y}_{t-1}, \ldots, \mathbf{W}\mathbf{Y}_{t-p}, \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_{t-p})$$

is a $N \times m$ matrix and $\mathbf{D}_t(\boldsymbol{\theta})$ is the $N \times N$ diagonal matrix with diagonal elements equal to $\lambda_{i,t}(\boldsymbol{\theta})$ for $i = 1, \ldots, N$. The Hessian matrix is given by

$$\mathbf{H}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{C}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \sum_{t=1}^{T} \mathbf{h}_{Nt}(\boldsymbol{\theta}) \,, \tag{4.12}$$

with $\mathbf{C}_t(\boldsymbol{\theta}) = \text{diag}\left\{ y_{1,t}/\lambda_{1,t}^2(\boldsymbol{\theta}) \ldots y_{N,t}/\lambda_{N,t}^2(\boldsymbol{\theta}) \right\}$ and the conditional information matrix is

$$\mathbf{B}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \sum_{t=1}^{T} \mathbf{b}_{Nt}(\boldsymbol{\theta}) \,,$$

where $\boldsymbol{\Sigma}_t(\boldsymbol{\theta}) = \mathrm{E}(\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T | \mathcal{F}_{t-1})$ denotes the *true* conditional covariance matrix of the vector $\mathbf{Y}_t$ and we have defined $\boldsymbol{\xi}_t \equiv \mathbf{Y}_t - \boldsymbol{\lambda}_t$. Expectation is taken with respect to the stationary distribution of $\{\mathbf{Y}_t\}$. We drop the dependence on $\boldsymbol{\theta}$ when a quantity is evaluated at $\boldsymbol{\theta}_0$.

Proposition 9. Consider model (4.2). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. Suppose that $\boldsymbol{\Theta}$ is compact and assume that the true value $\boldsymbol{\theta}_0$ belongs to the interior of $\boldsymbol{\Theta}$. Suppose that at the true value $\boldsymbol{\theta}_0$, the condition of Proposition 6 hold. Then, there exists a fixed open neighbourhood , say $O(\boldsymbol{\theta}) = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \delta\}$, of $\boldsymbol{\theta}_0$ such that with probability tending to 1 as $T \to \infty$, the equation $S_{NT}(\boldsymbol{\theta}) = 0$ has a unique solution, say $\tilde{\boldsymbol{\theta}}$. Moreover, $\tilde{\boldsymbol{\theta}}$ is consistent and asymptotically normal:

$$\sqrt{T}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{H}_N^{-1} \mathbf{B}_N \mathbf{H}_N^{-1}) \,,$$

with

$$\mathbf{H}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right], \tag{4.13}$$

$$\mathbf{B}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right]. \tag{4.14}$$

Proposition 9 follows immediately from Theorem 4.1 in Fokianos et al. (2020). Proposition 9 applies to the log-linear model (4.6), provided that $\mathrm{E}[\exp(r\,|\boldsymbol{\nu}_t|)] < \infty$, for any $r > 0$. Then, we have that the score function is given by:

$$\mathbf{S}_{NT}(\boldsymbol{\theta}) \quad = \sum_{t=1}^{T} \sum_{i=1}^{N} \Big(y_{i,t} - \exp(\nu_{i,t}(\boldsymbol{\theta})) \Big) \frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big(\mathbf{Y}_t - \exp(\boldsymbol{\nu}_t(\boldsymbol{\theta})) \Big), \tag{4.15}$$

where

$$\frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = (\mathbf{1}_N, \mathbf{W} \log(\mathbf{1}_N + \mathbf{Y}_{t-1}), \ldots, \mathbf{W} \log(\mathbf{1}_N + \mathbf{Y}_{t-p}), \log(\mathbf{1}_N + \mathbf{Y}_{t-1}), \ldots, \log(\mathbf{1}_N + \mathbf{Y}_{t-p}))$$

is a $N \times m$ matrix, and

$$\mathbf{H}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \,, \tag{4.16}$$

$$\mathbf{B}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \,,$$

where $\mathbf{D}_t(\boldsymbol{\theta})$ is the $N \times N$ diagonal matrix with diagonal elements equal to $\exp(\nu_{i,t}(\boldsymbol{\theta}))$ for $i = 1, \ldots, N$ and $\boldsymbol{\Sigma}_t(\boldsymbol{\theta}) = \mathrm{E}(\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T | \mathcal{F}_{t-1})$ with $\boldsymbol{\xi}_t = \mathbf{Y}_t - \exp(\boldsymbol{\nu}_t(\boldsymbol{\theta}))$. Moreover,

$$\mathbf{H}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right], \tag{4.17}$$

$$\mathbf{B}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right] \tag{4.18}$$

are respectively (minus) the Hessian matrix and the information matrix.

4.3.2 Quasi-likelihood inference for increasing N

Proposition 9 establishes asymptotic results when $T \to \infty$ and N fixed. In the paper $\mathbf{W}$ is a nonrandom sequence of matrices indexed by N. In this case, the specification of the asymptotic properties for $N \to \infty$ and $T \to \infty$ allows to establish a double-dimensional "spatio-temporal" type of consistency and asymptotic normality of the estimator. The results established in the previous section cannot be extended to such asymptotic regime because no ergodicity results are available, as $\min\{N, T\} \to \infty$. Moreover, the definition of stationarity for an N-dimensional time series $\mathbf{Y}_t \in \mathbb{R}^N$ when $N \to \infty$ does not seem to be generally established in the literature. Consequently, we propose here an alternative proof based on the previous stationarity results (with fixed N) and no ergodicity required. Define $l_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{i=1}^{N} l_{i,t}(\boldsymbol{\theta})$, where $l_{i,t}(\boldsymbol{\theta}) = y_{i,t}\log\lambda_{i,t}(\boldsymbol{\theta}) - \lambda_{i,t}(\boldsymbol{\theta})$. Let M be a finite constant.

Assumption 1. The following limits exist, at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$:

(i) $\lim_{N\to\infty} N^{-1}\mathbf{H}_N = \mathbf{H}$, with $\mathbf{H}$ a $m \times m$ positive definite matrix, where $\mathbf{H}_N$ is defined by (4.13).

(ii) $\lim_{N\to\infty} N^{-1}\mathbf{B}_N = \mathbf{B}$, with $\mathbf{B}$ a $m \times m$ positive definite matrix, where $\mathbf{B}_N$ is defined by (4.14).

(iii) Assume the third derivative of the quasi-log-likelihood (4.10) is bounded by functions m_{it} which satisfy $\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\mathrm{E}(m_{it}) = M$.

Assumption 2. For the linear model (4.4) assume

(i) $\frac{1}{N}\sum_{i,j=1}^{N}\|\xi_{it}\xi_{jt}\|_a < \infty$, for some $a \geq 4$.

(ii) The process $\left\{\boldsymbol{\xi}_t = \mathbf{Y}_t - \boldsymbol{\lambda}_t,\, \mathcal{F}_t^N : N \in \mathbb{N}, t \in \mathbb{Z}\right\}$ is α-mixing, $\mathcal{F}_t^N = \sigma\left(\xi_{is} : 1 \leq i \leq N, s \leq t\right)$.

Assumptions 1-(i) and 1-(ii) are type of law of large number assumptions, which are quite standard in the existing literature, since little is known about the behaviour of the distribution as $N \to \infty$. See assumption C3 of Zhu et al. (2017) and assumption C2.3 of Zhu et al. (2019). To clarify this, set $p = 1$, so $m = 3$. Define $\tilde{Y}_{it-1} = w_i^T\mathbf{Y}_{t-1}$ and $\sigma_{ijt} = \mathrm{E}(\xi_{it}\xi_{jt}|\mathcal{F}_{Nt-1})$. Then, the matrices $\mathbf{H}_N$, $\mathbf{B}_N$ in (4.13, 4.14), evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, will be

$$\mathbf{H}_N = \begin{pmatrix} \mathbf{H}_N^{(11)} & \mathbf{H}_N^{(12)} & \mathbf{H}_N^{(13)} \\ & \mathbf{H}_N^{(22)} & \mathbf{H}_N^{(23)} \\ & & \mathbf{H}_N^{(33)} \end{pmatrix}, \quad \mathbf{B}_N = \begin{pmatrix} \mathbf{B}_N^{(11)} & \mathbf{B}_N^{(12)} & \mathbf{B}_N^{(13)} \\ & \mathbf{B}_N^{(22)} & \mathbf{B}_N^{(23)} \\ & & \mathbf{B}_N^{(33)} \end{pmatrix},$$

where

$$\mathbf{H}_N^{(11)} = \mathrm{E}\left[\mathbf{1}_N^T\mathbf{D}_t^{-1}\mathbf{1}_N\right] = \sum_{i=1}^{N}\mathrm{E}\left(\frac{1}{\lambda_{i,t}}\right), \quad \mathbf{H}_N^{(12)} = \mathrm{E}\left[\mathbf{1}_N^T\mathbf{D}_t^{-1}\mathbf{W}\mathbf{Y}_{t-1}\right] = \sum_{i=1}^{N}\mathrm{E}\left(\frac{\tilde{Y}_{it-1}}{\lambda_{i,t}}\right),$$

$$\mathbf{H}_N^{(13)} = \mathrm{E}\left[\mathbf{1}_N^T\mathbf{D}_t^{-1}\mathbf{Y}_{t-1}\right] = \sum_{i=1}^{N}\mathrm{E}\left(\frac{Y_{it-1}}{\lambda_{i,t}}\right), \quad \mathbf{H}_N^{(22)} = \mathrm{E}\left[\mathbf{Y}_{t-1}^T\mathbf{W}^T\mathbf{D}_t^{-1}\mathbf{W}\mathbf{Y}_{t-1}\right] = \sum_{i=1}^{N}\mathrm{E}\left(\frac{\tilde{Y}_{it-1}^2}{\lambda_{i,t}}\right),$$

$$\mathbf{H}_N^{(23)} = \mathrm{E}\left[\mathbf{Y}_{t-1}^T\mathbf{W}^T\mathbf{D}_t^{-1}\mathbf{Y}_{t-1}\right] = \sum_{i=1}^{N}\mathrm{E}\left(\frac{\tilde{Y}_{it-1}Y_{it-1}}{\lambda_{i,t}}\right), \quad \mathbf{H}_N^{(33)} = \mathrm{E}\left[\mathbf{Y}_{t-1}^T\mathbf{D}_t^{-1}\mathbf{Y}_{t-1}\right] = \sum_{i=1}^{N}\mathrm{E}\left(\frac{Y_{it-1}^2}{\lambda_{i,t}}\right),$$

and

$$\mathbf{B}_N^{(11)} = \mathrm{E}\left[\mathbf{1}_N^T \mathbf{D}_t^{-1} \mathbf{\Sigma}_t \mathbf{D}_t^{-1} \mathbf{1}_N\right] = \sum_{i,j=1}^{N} \mathrm{E}\left(\frac{\xi_{it}\xi_{jt}}{\lambda_{it}\lambda_{jt}}\right),$$

$$\mathbf{B}_N^{(12)} = \mathrm{E}\left[\mathbf{1}_N^T \mathbf{D}_t^{-1} \mathbf{\Sigma}_t \mathbf{D}_t^{-1} \mathbf{W}\mathbf{Y}_{t-1}\right] = \sum_{i,j=1}^{N} \mathrm{E}\left(\frac{\sigma_{ijt}\tilde{Y}_{it-1}}{\lambda_{it}\lambda_{jt}}\right),$$

$$\mathbf{B}_N^{(13)} = \mathrm{E}\left[\mathbf{1}_N^T \mathbf{D}_t^{-1} \mathbf{\Sigma}_t \mathbf{D}_t^{-1} \mathbf{Y}_{t-1}\right] = \sum_{i,j=1}^{N} \mathrm{E}\left(\frac{\sigma_{ijt}Y_{it-1}}{\lambda_{it}\lambda_{jt}}\right),$$

$$\mathbf{B}_N^{(22)} = \mathrm{E}\left[\mathbf{Y}_{t-1}^T \mathbf{W}^T \mathbf{D}_t^{-1} \mathbf{\Sigma}_t \mathbf{D}_t^{-1} \mathbf{W}\mathbf{Y}_{t-1}\right] = \sum_{i,j=1}^{N} \mathrm{E}\left(\frac{\sigma_{ijt}\tilde{Y}_{it-1}\tilde{Y}_{jt-1}}{\lambda_{it}\lambda_{jt}}\right),$$

$$\mathbf{H}_N^{(23)} = \mathrm{E}\left[\mathbf{Y}_{t-1}^T \mathbf{W}^T \mathbf{D}_t^{-1} \mathbf{\Sigma}_t \mathbf{D}_t^{-1} \mathbf{Y}_{t-1}\right] = \sum_{i,j=1}^{N} \mathrm{E}\left(\frac{\sigma_{ijt}\tilde{Y}_{it-1}Y_{jt-1}}{\lambda_{it}\lambda_{jt}}\right),$$

$$\mathbf{H}_N^{(33)} = \mathrm{E}\left[\mathbf{Y}_{t-1}^T \mathbf{D}_t^{-1} \mathbf{\Sigma}_t \mathbf{D}_t^{-1} \mathbf{Y}_{t-1}\right] = \sum_{i,j=1}^{N} \mathrm{E}\left(\frac{\sigma_{ijt}Y_{it-1}Y_{jt-1}}{\lambda_{it}\lambda_{jt}}\right),$$

Assumption 1-(i) requires the laws of large number $\lim_{N\to\infty} N^{-1}\mathbf{H}_N^{(k,l)} = h_{kl}$, $\lim_{N\to\infty} N^{-1}\mathbf{B}_N^{(k,l)} = b_{kl}$, where $h_{k,l}$ and $b_{k,l}$ are constants, for $k, l = 1, 2, 3$ and $(k, l) = (l, k)$.

In the setup we study, however, we require two "regularity" conditions since under the quasi-likelihood inference the information matrix and the Hessian matrix are in general different. This is not the case in Zhu et al. (2017), since these authors consider least squares regression under i.i.d. assumption of the error terms. For the same reason, a condition on the derivative is usually required for the quasi-likelihood approach, as in Assumption 1-(iii).

The condition Assumption 2-(i) can also be seen as a law of large numbers-type of assumption which is additional in our case, since the error term does not consist of an i.i.d. sequence. Moreover, for the result of Fokianos et al. (2020, Prop. 3.1-3.4), Assumption 2-(i) is satisfied for fixed N; we conjecture that this still holds true when N increases, as in this case the behaviour of the distribution of the process is unknown. This kind of assumption is common in the literature of high-dimensional processes, see, for example, Assumption M1 in Stock and Watson (2002).

Finally, Assumption 2-*(ii)* is a crucial assumption we adopt as we study processes with dependent errors (see Doukhan (1994) for definition of α-mixing). The α-mixing is a measure of *asymptotic independence* of the process and it is weaker than the i.i.d. assumption made by Zhu et al. (2017, 2019). In particular, the process defined in Assumption 2-*(ii)* is an α-mixing array, namely,

$$\alpha(J) = \sup_{t\in\mathbb{Z}, N\geq 1} \sup_{A\in\mathcal{F}_{-\infty,t}^N, B\in\mathcal{F}_{t+J,\infty}^N} |\mathrm{P}(A\cap B) - \mathrm{P}(A)\mathrm{P}(B)| \xrightarrow{m\to\infty} 0$$

where $\mathcal{F}_t^N \equiv \mathcal{F}_{-\infty,t}^N = \sigma\left(\xi_{is} : 1 \leq i \leq N, s \leq t\right)$, $\mathcal{F}_{t+J,\infty}^N = \sigma\left(\xi_{is} : 1 \leq i \leq N, s \geq t+J\right)$ and it is clear that the dependence between two events A and B tends to vanish as they are spaced in time, uniformly in N. Moreover, note that no rate of decay for the dependence measured by $\alpha(J)$ along time is specified, as a consequence, the α-mixing process can depends on several lags of its past before becoming "asymptotically" independent. When N is fixed and $p = 1$, by Fokianos et al. (2020, Prop. 3.1-3.4), the assumptions $|||\mathbf{G}|||_1 < 1$ or $|||\mathbf{G}|||_2 < 1$ are sufficient conditions for obtaining an α-mixing process $\{\boldsymbol{\xi}_t : t \in \mathbb{Z}\}$.

Note that we develop an approach where no further assumptions on the network structure are required, compare with Zhu et al. (2017, 2019, Ass. C2.1-C2.2). This leads to a more flexible framework for modelling network processes. Following the discussion in Zhu et al. (2019, p. 351), assumption C2.2 in Zhu et al. (2017, 2019) might not hold true when there exists considerable heterogeneity among nodes of the network (e.g., a social network with few "superstars" and several low-active nodes). Such an assumption, though, is not required by our approach.

Lemma 7. For the linear model (4.4), suppose the condition of Proposition 6 and Assumptions 1-2 hold. Consider $\mathbf{S}_{NT}$ and $\mathbf{H}_{NT}$ defined as in (4.11) and (4.12), respectively. Then, as $\min\{N,T\} \to \infty$

1. $(NT)^{-1}\mathbf{H}_{NT} \xrightarrow{p} \mathbf{H}$,
2. $(NT)^{-\frac{1}{2}}\mathbf{S}_{NT} \xrightarrow{d} N(0,\mathbf{B})$,
3. $\max_{j,l,k} \sup_{\boldsymbol{\theta}\in\mathcal{O}(\boldsymbol{\theta}_0)} \left| \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_l\partial\boldsymbol{\theta}_k} \right| \leq M_{NT} \xrightarrow{p} M$,

where $M_{NT} \coloneqq \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} m_{i,t}$. The proof is postponed to the Appendix.

Theorem 15. Consider model (4.4). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}_+^m$. Suppose that $\boldsymbol{\Theta}$ is compact and assume that the true value $\boldsymbol{\theta}_0$ belongs to the interior of $\boldsymbol{\Theta}$. Suppose that the conditions of Lemma 7 hold. Then, there exists a fixed open neighbourhood $\mathcal{O}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \delta\}$ of $\boldsymbol{\theta}_0$ such that with probability tending to 1 as $\min\{N,T\} \to \infty$, for the score function (4.11), the equation $S_{NT}(\boldsymbol{\theta}) = 0$ has a unique solution, called $\hat{\boldsymbol{\theta}}$, which is consistent and asymptotically normal:

$$\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}).$$

Lemma 7 and Taniguchi and Kakizawa (2000, Thm. 3.2.23) adapted to double-indexed convergence, for instance, guarantees the conclusion of Theorem 15.

We now state the analogous result for the log-linear model (4.8) and the notation corresponds to eq. (4.15)–(4.18).

Assumption 1′. Assume the same conditions as in Assumption 1 but with $\mathbf{H}_N$ and $\mathbf{B}_N$ defined in (4.17) and (4.18), respectively.

Assumption 2′. For the log-linear model (4.8) assume

(i) $\frac{1}{N}\sum_{i,j=1}^{N} \|\xi_{it}\xi_{jt}\|_a < \infty$, $\max_{i\geq 1} \mathrm{E}|Y_{it}|^r < \infty$, $\max_{i\geq 1} \mathrm{E}[\exp(r|\nu_{it}|)] < \infty$, for any $r \geq 1$ and some $a \geq 4$.

(ii) $\{\boldsymbol{\psi}_t = \log(\mathbf{1} + \mathbf{Y}_t) - \boldsymbol{\nu}_t, \mathcal{F}_t^N : N \in \mathbb{N}, t \in \mathbb{Z}\}$ is α-mixing; $\mathcal{F}_t^N = \sigma(\psi_{is} : 1 \leq i \leq N, s \leq t)$.

The same discussion about Assumptions 1′ and 2′ applies similar to the QMLE of the linear model. The existence of exponential moments is of crucial importance to study the properties of log-linear models. see Fokianos and Tjøstheim (2011) and Fokianos et al. (2020), among others.

Lemma 8. Let $\mathbf{S}_{NT}$ and $\mathbf{H}_{NT}$ as in (4.15) and (4.16). Then, for the log-linear model (4.8), under the condition of Proposition 8 and Assumptions 1′-2′ the conclusion of Lemma 7 holds.

The proof is postponed to the Appendix.

Theorem 16. Consider model (4.8). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. Suppose that $\boldsymbol{\Theta}$ is compact and assume that the true value $\boldsymbol{\theta}_0$ belongs to the interior of $\boldsymbol{\Theta}$. Suppose that the conditions of Lemma 8 hold. Then, there exists a fixed open neighbourhood $\mathcal{O}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \delta\}$ of $\boldsymbol{\theta}_0$ such that with probability tending to 1 as $\min\{N,T\} \to \infty$, for the score function (4.15), the equation $S_{NT}(\boldsymbol{\theta}) = 0$ has a unique solution, called $\hat{\boldsymbol{\theta}}$, which is consistent and asymptotically normal:

$$\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}).$$

The conclusion follows as above.

In practical application one needs to specify a suitable estimator for the limiting covariance matrix of the quasi maximum likelihood estimators. To this aim define the following matrix

$$\hat{\mathbf{B}}_{NT}(\hat{\boldsymbol{\theta}}) = \sum_{t=1}^{T} \mathbf{s}_{Nt}(\hat{\boldsymbol{\theta}}) \mathbf{s}_{Nt}(\hat{\boldsymbol{\theta}})^T .$$

Let $\mathbf{V} \coloneqq \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ and $\mathbf{V}(\hat{\boldsymbol{\theta}}) \coloneqq (NT)\mathbf{H}_{NT}^{-1}(\hat{\boldsymbol{\theta}})\hat{\mathbf{B}}_{NT}(\hat{\boldsymbol{\theta}})\mathbf{H}_{NT}^{-1}(\hat{\boldsymbol{\theta}})$. The following results establish the inference for the limiting covariance matrix of Theorems 15 and 16, respectively.

Theorem 17. Consider model (4.4). Suppose the conditions of Theorem 15 hold true. Moreover, assume that $N^{-1}\sum_{i,j=1}^{N} \|Y_{it}Y_{jt}\|_2 < \infty$. Then, as $\min\{N,T\} \to \infty$, $\mathbf{V}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{V}$.

The proof is postponed to the Appendix.

Theorem 18. Consider model (4.8). Suppose the conditions of Theorem 16 hold true. Moreover, assume that $N^{-1}\sum_{i,j=1}^{N} \|\exp(\nu_{it})\exp(\nu_{jt})\|_2 < \infty$. Then, as $\min\{N,T\} \to \infty$, $\mathbf{V}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{V}$.

The proof is analogous to the proof of Theorem 17, therefore is omitted.

4.4 Applications

4.4.1 Simulations

We study finite sample behaviour of the QMLE for models (4.2) and (4.6). For this goal we ran a simulation study with $S = 1000$ repetitions and different time series length and network dimension. We consider the cases $p = 1$ and 2. The adjacency matrix is generated by the lag-one Stochastic Block model ($K = 5$ blocks) using $(\beta_0, \beta_1, \beta_2)^T = (0.2, 0.4, 0.5)^T$. The observed time series are generated using the copula-based data generating process of Fokianos et al. (2020). The network density is set equal to 0.3. We performed simulations with a network density equal to 0.5, as well, but we obtained similar results, hence we do not reported these. Tables4.1 and 4.2 summarize the simulation results. Additional findings are given in the Appendix–see Tables C-1–C-6.

The estimates for parameters and their standard errors (in brackets) are obtained by averaging out the results from all simulations; the third row below each coefficient shows the percentage frequency of t-tests which reject $H_0 : \beta = 0$ at the level 1% over the S simulations. We also report the percentage of cases where various information criteria select the correct generating model. In this study, we employ the Akaike (AIC), the Bayesian (BIC) and the Quasi (QIC) information criteria. The latter is a special case of the AIC which takes into account that estimation is done by quasi-likelihood methods. See Pan (2001) for more details.

We observe that when there is strong correlation between count variables $Y_{i,t}$–see Table 4.1– and T is small when compared to the network size N, then the estimates are biased. The same conclusion is drawn from Table C-1. Instead, when both T and N are reasonably large (or at least T is large), then the estimates are close to the real values and the standard errors are small. Standard errors reduce as T increases–this should be expected. Regarding estimators of the log-linear model (see Table 4.2 and C-4), we obtain the same conclusions. Note that the approximations for network ($\hat{\beta}_1$) and lagged ($\hat{\beta}_2$) effects is better when compared to the approximation of intercept ($\hat{\beta}_0$).

The t-tests and percentage of right selections due to various information criteria provide empirical confirmation for the model selection procedure. Again, we note that when T is small then there is no definite winner among all of them. Based on these results, the QIC provides the best selection procedure for the case of the linear model; its success selection rate is about 94%. The BIC shows better performance only when N is small and this is so because

it tends to select models with fewer parameters. The same conclusions are reached for the case of the log-linear model, even though the rate of right selections for the QIC does not exceed 87%. However, the QIC is more robust, especially when used for misspecified models.

To validate these results, we consider the case where all series are independent (Gaussian copula with $\rho = 0$). Then QMLE provides satisfactory results if N is large enough, even if T is small (see Table C-2, C-5). When $\rho > 0$, both the temporal size T and the network size N are required to be reasonably large in order to obtain good inferential results. From the QQ-plot shown in Figure 4.4 we can conclude that, with N and T large enough, the asserted asymptotic normality is quite adequate. For this plot, the data were generated by a linear model with a Gaussian copula ($\rho = 0.5$) and $N = 100$. A more extensive discussion and further simulation results can be found in the Appendix.

Table 4.1: Estimators obtained from $S = 1000$ simulations of model (4.2), for various values of N and T. Data are generated by using the Gaussian copula with $\rho = 0.5$ and $p = 1$. Model (4.2) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	*AIC*	*BIC*	*QIC*
20	100	0.202	0.395	0.492	0.199	0.384	0.485	0.017	0.009	87.0	97.1	93.7
		(0.029)	(0.046)	(0.041)	(0.030)	(0.056)	(0.046)	(0.050)	(0.029)			
		100	100	100	100	99.9	100	0.4	0.1			
	200	0.202	0.396	0.496	0.199	0.389	0.491	0.011	0.006	89.7	97.3	94.2
		(0.021)	(0.032)	(0.030)	(0.021)	(0.039)	(0.032)	(0.034)	(0.020)			
		100	100	100	100	100	100	0.1	0.5			
100	10	0.254	0.337	0.438	0.240	0.316	0.424	0.039	0.016	78.9	79.9	85.1
		(0.104)	(0.077)	(0.079)	(0.103)	(0.099)	(0.095)	(0.109)	(0.071)			
		13.7	69	85.7	5.5	36.2	64.8	0.6	0.1			
	20	0.235	0.366	0.465	0.227	0.351	0.454	0.025	0.011	77.6	81.2	90.7
		(0.075)	(0.057)	(0.059)	(0.076)	(0.074)	(0.069)	(0.072)	(0.044)			
		65.3	96.3	99.3	56	90.4	98.6	0.8	0.3			
	100	0.207	0.393	0.491	0.204	0.385	0.486	0.011	0.005	75.0	83.8	93.6
		(0.033)	(0.025)	(0.026)	(0.034)	(0.034)	(0.031)	(0.031)	(0.018)			
		100	100	100	100	100	100	0.4	0.1			
	200	0.202	0.396	0.496	0.200	0.390	0.492	0.008	0.004	72.1	83.1	94.1
		(0.023)	(0.018)	(0.019)	(0.024)	(0.024)	(0.022)	(0.022)	(0.013)			
		100	100	100	100	100	100	0.3	0.2			

Table 4.2: Estimators obtained from $S = 1000$ simulations of model (4.6), for various values of N and T. Data are generated by using the Gaussian copula with $\rho = 0.5$ and $p = 1$. Model (4.6) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	*AIC*	*BIC*	*QIC*
20	100	0.209	0.402	0.492	0.212	0.401	0.494	0.003	-0.006	60.4	85.5	82.5
		(0.069)	(0.022)	(0.039)	(0.074)	(0.038)	(0.048)	(0.043)	(0.040)			
		64.8	100	100	57.7	100	100	0.6	0.4			
	200	0.204	0.403	0.494	0.206	0.402	0.495	0.003	-0.003	61.6	90.0	84.9
		(0.049)	(0.016)	(0.027)	(0.053)	(0.027)	(0.034)	(0.031)	(0.028)			
		93.2	100	100	89.2	100	100	0.6	0.3			
100	10	0.299	0.392	0.443	0.301	0.368	0.443	0.039	-0.011	30.2	35.5	58.2
		(0.195)	(0.043)	(0.078)	(0.191)	(0.077)	(0.087)	(0.088)	(0.069)			
		12.4	99.4	87.8	10.9	67.5	72.0	1.3	0.6			
	20	0.265	0.398	0.465	0.269	0.390	0.472	0.015	-0.015	25.6	34.1	69.8
		(0.145)	(0.028)	(0.056)	(0.146)	(0.062)	(0.069)	(0.071)	(0.053)			
		20.5	100	99.8	20.5	99.2	99.4	1.7	0.6			
	100	0.216	0.401	0.492	0.218	0.402	0.496	0.000	-0.006	23.3	44.3	82.3
		(0.065)	(0.012)	(0.025)	(0.068)	(0.030)	(0.033)	(0.035)	(0.026)			
		74.2	100	100	70.9	100	100	0.7	0.5			
	200	0.209	0.401	0.495	0.210	0.399	0.496	0.002	-0.002	26.6	51.0	86.9
		(0.046)	(0.008)	(0.018)	(0.048)	(0.022)	(0.022)	(0.025)	(0.018)			
		96.7	100	100	95.6	100	100	0.5	0.2			

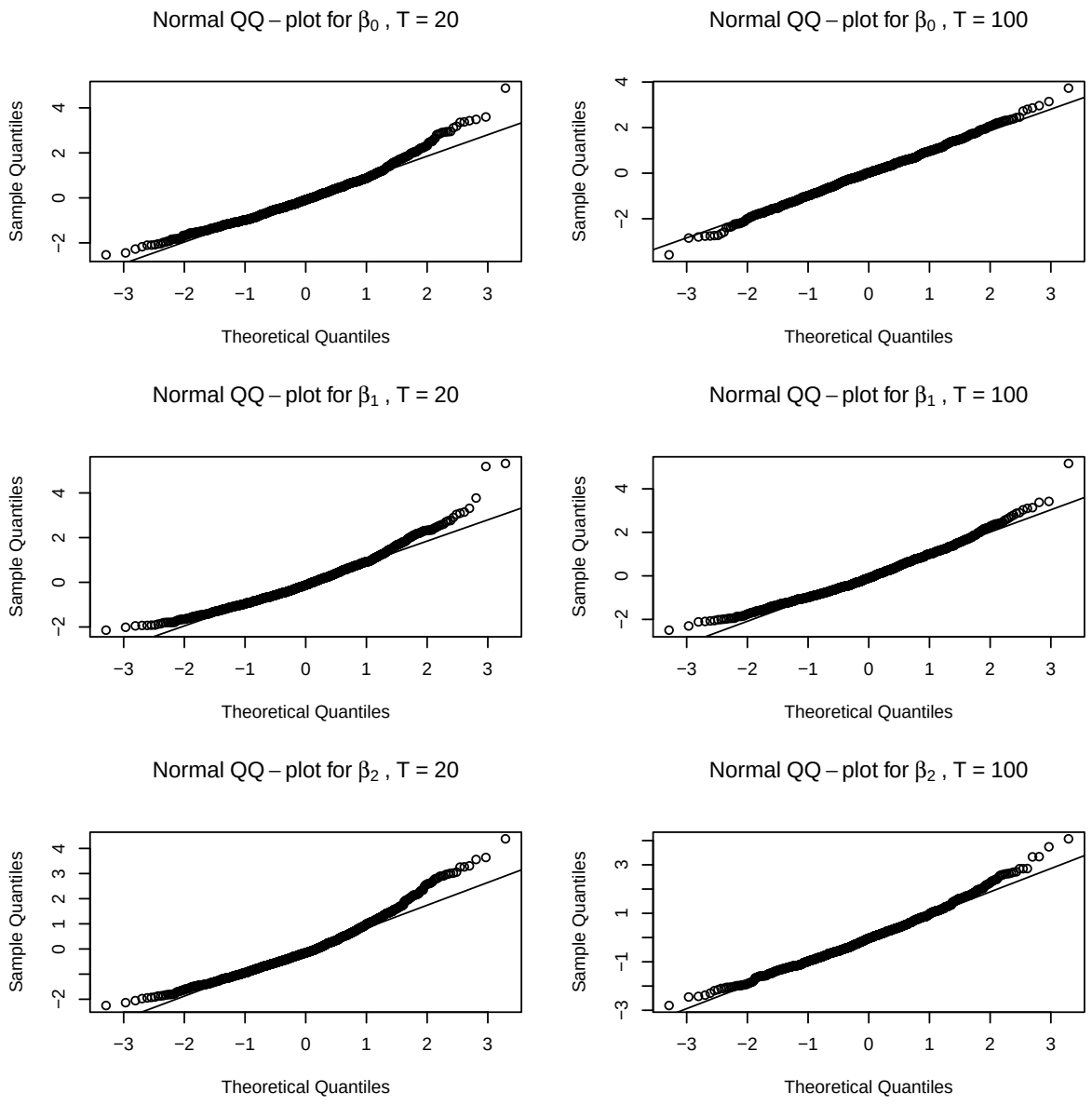

Figure 4.4: QQ-plots for the linear model, Gaussian copula with $\rho = 0.5$, $N = 100$. Left: $T = 20$. Right: $T = 100$.

4.4.2 Data analysis

The application on real data concerns the monthly number of burglaries on the south side of Chicago from 2010-2015 ($T = 72$). The counts are registered for the $N = 552$ census block groups. The data are taken by Clark et al. (2018), `https://github.com/nick3703/Chicago-Data`. The undirected network structure raises naturally, as an edge between block i and j is set if the locations share a border. The density of the network is 1.74%. The maximum number of burglaries in a month in a census block is 17. The variance to mean ratio in the data is 1.82, suggesting there is some overdispersion in the data. The median of degrees is 5. On this dataset we fit the linear and log-linear PNAR(1) and PNAR(2) model. The results are summarized in Table 4.3-4.4. All the models have

significant parameters. The magnitude of the network effects β_{11} and β_{12} seems reasonable, as an increasing number of burglaries in a block can lead to a growth in the same type of crime committed in a close area. Also, the lagged effects have an increasing impact on the counts. Interestingly, the log-linear model is able to account for the general downward trend registered from 2010 to 2015 for this type of crime in the area analysed. All the information criteria select the PNAR(2) models, in accordance with the significance of the estimates.

Table 4.3: Estimation results for Chicago crime data.

	Linear PNAR(1)			Log-linear PNAR(1)		
	Estimate	SE ($\times 10^2$)	p-value	Estimate	SE ($\times 10^2$)	p-value
β_0	0.4551	2.1607	<0.01	-0.5158	3.8461	<0.01
β_1	0.3215	1.2544	<0.01	0.4963	2.8952	<0.01
β_2	0.2836	0.8224	<0.01	0.5027	1.2105	<0.01
	Linear PNAR(2)			Log-linear PNAR(2)		
	Estimate	SE ($\times 10^2$)	p-value	Estimate	SE ($\times 10^2$)	p-value
β_0	0.3209	1.8931	<0.01	-0.5059	4.7605	<0.01
β_{11}	0.2076	1.1742	<0.01	0.2384	3.4711	<0.01
β_{21}	0.2287	0.7408	<0.01	0.3906	1.2892	<0.01
β_{12}	0.1191	1.4712	<0.01	0.0969	3.3404	<0.01
β_{22}	0.1626	0.7654	<0.01	0.2731	1.2465	<0.01

Table 4.4: Information criteria for Chicago crime data. Smaller values in bold.

	AIC$\times 10^{-3}$		BIC$\times 10^{-3}$		QIC$\times 10^{-3}$	
	linear	log-linear	linear	log-linear	linear	log-linear
PNAR(1)	115.06	115.37	115.07	115.38	115.11	115.44
PNAR(2)	**111.70**	**112.58**	**111.72**	**112.60**	**111.76**	**112.68**

Acknowledgements

Both authors appreciate the hospitality of the Department of Mathematics & Statistics at Lancaster University, where this work was initiated. This work has been funded by the European Regional development Fund and the Republic of Cyprus through the Research and innovation Foundation, under the project INFRASTRUCTURES/1216/0017 (IRIDA).

Appendix

Moments for the linear PNAR(p) model

It is easy to derive some elementary properties of the linear NAR(p) model. Fix $\boldsymbol{\mu} = (\mathbf{I}_N - (\mathbf{G}_1 + \cdots + \mathbf{G}_P))^{-1}\boldsymbol{\beta}_0$; we can again rewrite model (4.3) as a Vector Autoregressive VAR(1) model

$$\mathbf{Y}_t - \boldsymbol{\mu} = \mathbf{G}_1(\mathbf{Y}_{t-1} - \boldsymbol{\mu}) + \cdots + \mathbf{G}_p(\mathbf{Y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{\xi}_t\,,$$

where $\boldsymbol{\xi}_t$ is a martingale difference sequence, and rearrange it in a Np-dimensional VAR(1) form by

$$\mathbf{Y}^*_t - \boldsymbol{\mu}^* = \mathbf{G}^*(\mathbf{Y}^*_{t-1} - \boldsymbol{\mu}^*) + \boldsymbol{\Xi}_t\,. \tag{C-1}$$

Here we have $\mathbf{Y}_t^* = (\mathbf{Y}_t^T, \mathbf{Y}_{t-1}^T, \ldots, \mathbf{Y}_{t-p+1}^T)^T$, $\boldsymbol{\mu}^* = (\mathbf{I}_{Np} - \mathbf{G}^*)^{-1}\mathbf{B}_0$, $\mathbf{B}_0 = (\boldsymbol{\beta}_0^T, \mathbf{0}_{N(p-1)}^T)^T$ $\boldsymbol{\Xi}_t = (\boldsymbol{\xi}_t, \mathbf{0}_{N(p-1)}^T)^T$, where $\mathbf{0}_{N(p-1)}$ is a $N(p-1) \times 1$ vector of zeros, and

$$\mathbf{G}^* = \begin{pmatrix} \mathbf{G}_1 & \mathbf{G}_2 & \cdots & \mathbf{G}_{p-1} & \mathbf{G}_p \\ \mathbf{I}_N & \mathbf{0}_{N,N} & \cdots & \mathbf{0}_{N,N} & \mathbf{0}_{N,N} \\ \mathbf{0}_{N,N} & \mathbf{I}_N & \cdots & \mathbf{0}_{N,N} & \mathbf{0}_{N,N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_{N,N} & \mathbf{0}_{N,N} & \cdots & \mathbf{I}_N & \mathbf{0}_{N,N} \end{pmatrix},$$

where $\mathbf{0}_{N,N}$ is a $N \times N$ matrix of zeros.

For model (C-1) we can find the unconditional mean $\mathrm{E}(\mathbf{Y}_t^*) = \boldsymbol{\mu}^*$ and variance $\mathrm{vec}[\mathrm{Var}(\mathbf{Y}_t^*)] = (\mathbf{I}_{(Np)^2} - \mathbf{G}^* \otimes \mathbf{G}^*)^{-1}\mathrm{vec}[\mathrm{E}(\boldsymbol{\Sigma}_t^*)]$ with $\mathrm{E}(\boldsymbol{\Sigma}_t^*) = \mathrm{E}(\boldsymbol{\Xi}_t\boldsymbol{\Xi}_t^T)$. For details about the VAR(1) representation of a VAR(p) model and its moments, see Lütkepohl (2005). Define the selection matrix $\mathbf{J} = (\mathbf{I}_N : \mathbf{0}_{N,N} : \cdots : \mathbf{0}_{N,N})$ with dimension $N \times Np$.

Proposition 10. Assume that N is fixed and $\sum_{h=1}^{p}(\beta_{1h} + \beta_{2h}) < 1$ in model (4.3). Then, model (4.4) has the following unconditional moments:

$$\begin{aligned} \mathrm{E}(\mathbf{Y}_t) &= \mathbf{J}\boldsymbol{\mu}^* = (\mathbf{I}_N - (\mathbf{G}_1 + \cdots + \mathbf{G}_P))^{-1}\boldsymbol{\beta}_0 = \boldsymbol{\mu}\,, \\ \mathrm{vec}[\mathrm{Var}(\mathbf{Y}_t)] &= (\mathbf{J} \otimes \mathbf{J})\mathrm{vec}[\mathrm{Var}(\mathbf{Y}_t^*)]\,, \\ \mathrm{vec}[\mathrm{Cov}(\mathbf{Y}_t, \mathbf{Y}_{t-h})] &= (\mathbf{J} \otimes \mathbf{J})(\mathbf{I}_{Np} - \mathbf{G}^*)^h\mathrm{vec}[\mathrm{Var}(\mathbf{Y}_t^*)]\,. \end{aligned}$$

Applying these results to model (4.1) (equivalently (4.2)), we obtain

$$\begin{aligned} \mathrm{E}(\mathbf{Y}_t) &= (\mathbf{I}_N - \mathbf{G})^{-1}\boldsymbol{\beta}_0 = \beta_0(1 - \beta_1 - \beta_2)^{-1}\mathbf{1}\,, \\ \mathrm{vec}[\mathrm{Var}(\mathbf{Y}_t)] &= (\mathbf{I}_{N^2} - \mathbf{G} \otimes \mathbf{G})^{-1}\mathrm{vec}[\mathrm{E}(\boldsymbol{\Sigma}_t)]\,, \\ \mathrm{vec}[\mathrm{Cov}(\mathbf{Y}_t, \mathbf{Y}_{t-h})] &= (\mathbf{I}_N - \mathbf{G})^h\mathrm{vec}[\mathrm{Var}(\mathbf{Y}_t)]\,. \end{aligned} \tag{C-2}$$

The mean of $\mathbf{Y}_t$ depends on the network effect β_1, the momentum effect β_2 and the structure of the network (via $\mathbf{W}$). The same fact holds for second moments structure; in addition, the conditional covariance $\boldsymbol{\Sigma}_t$ makes explicit the dependence on the copula correlation structure. We can observe that equations (C-2) are analogous to equations (2.4) and (2.5) of Zhu et al. (2017, Prop. 1), who analysed the continuous variable case. Then, the interpretations (Case 1 and 2 pag.1099-1100) and the potential applications (Section 3, pag.1105) apply also here for integer-valued case.

Empirical properties of the log-linear PNAR(1) model

We give here some insight on the structure of the model (4.6) above for the linear model. Here an explicit formulation of the unconditional moments is not possible. We report the sample statistics to estimate the unknown quantities and replicate the same baseline characteristics and the same scenarios of the linear case. In Figure C-1 we can see that, analogously to the linear case, the correlations among counts grow when more activity in the network is showed. However, here a more sparse matrix seems to slightly affect correlations. The general levels of correlations are higher than the linear case in Figure 4.1. The mean ranges around 1.7 and 2; it tends to rise with higher network activities up to 2.2. For the variance we find analogous results.

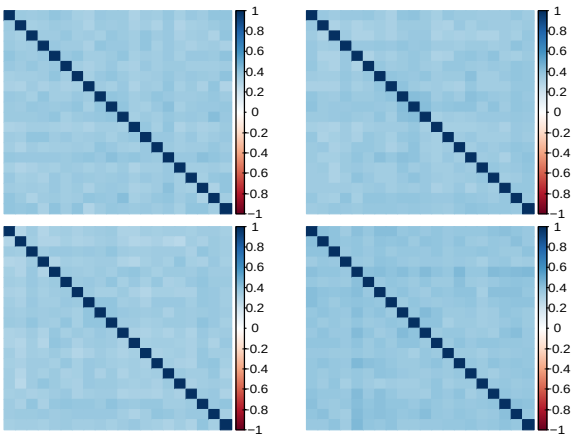

Figure C-1: Correlation matrix of model (4.5). Top-left: Data are generated by employing a stochastic block model with $K = 5$ and an adjacency matrix $\mathbf{A}$ with elements generated by $\mathrm{P}(a_{ij} = 1) = 0.3N^{-0.3}$, if i and j belong to the same block, and $\mathrm{P}(a_{ij} = 1) = 0.3N^{-1}$, otherwise. In addition, we employ a Gaussian copula with parameter $\rho = 0.5$, $(\beta_0, \beta_1, \beta_2) = (0.2, 0.1, 0.4)^T$, $T = 2000$ and $N = 20$. Top-right plot: Data are generated by employing a stochastic block model with $K = 5$ and an adjacency matrix $\mathbf{A}$ with elements generated by $\mathrm{P}(a_{ij} = 1) = 0.7N^{-0.0003}$ if i and j belong to the same block, and $\mathrm{P}(a_{ij} = 1) = 0.6N^{-0.3}$ otherwise. Same values for β's, T, N and choice of copula. Bottom-left: The same graph, as in the upper-left side but with $K = 10$. Bottom-right: The same graph, as in upper-right side but with $K = 10$.

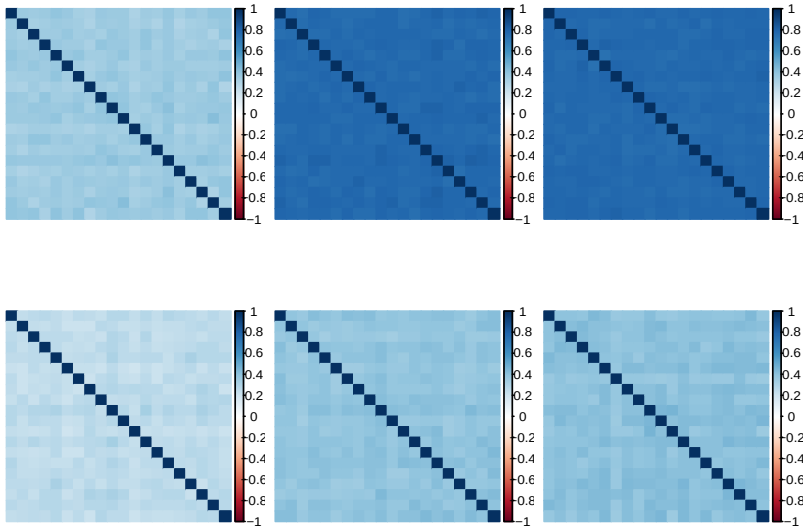

Figure C-2: Correlation matrix of model (4.5). Top: Data have been generated as in top-left of Figure C-1 (left), with copula correlation parameter $\rho = 0.9$ (middle) and as in the top-right of Figure C-1 but with copula parameter $\rho = 0.9$ (right). Bottom: same information as the top plot but data are generated by using a Clayton copula.

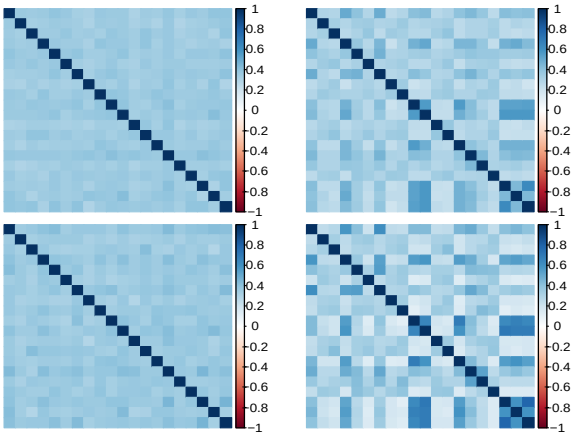

Figure C-3: Correlation matrix of model (4.5). Data have been generated as in top-left of Figure C-1 (top-left), higher network effect $\beta_1 = 0.4$ (top-right), higher momentum effect $\beta_2 = 0.6$ (lower-left) and higher network and momentum effect $\beta_1 = 0.3$, $\beta_2 = 0.6$ (lower-right).

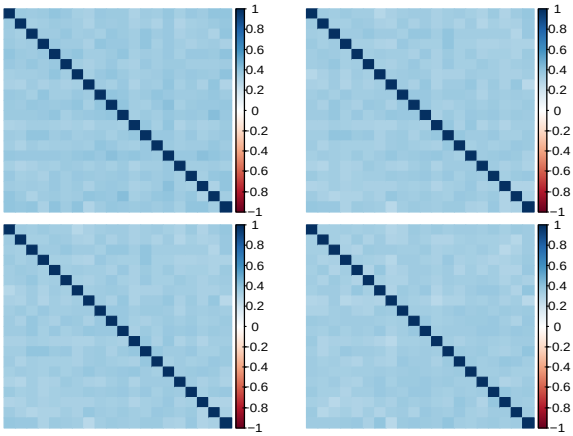

Figure C-4: Correlation matrix of model (4.5). Data have been generated as in top-left of Figure C-1 (top-left), negative network effect $\beta_1 = -0.1$ (top-right), negative momentum effect $\beta_2 = -0.4$ (lower-left) and negative network and momentum effect $\beta_1 = -0.1$, $\beta_2 = -0.4$ (lower-right).

Figure C-2 shows the outcomes obtained by varying the copula structure and the copula parameter ρ. The results are similar to Figure 4.2 but here the correlations tend to be more homogeneous. By adding positive weights to the network and momentum effect in Figure C-3 we notice comparable results with those of the linear model in Figure 4.3, but here the growth in parameters leads to a less severe effect on correlations. Significant increases in mean and variance are detected. In the log-linear model negative values for the parameters are allowed. In Figure C-4 we see no remarkable impact of negative coefficients on correlations. However, the sample means and variances decrease when compared to the corresponding plots produced using $\beta_1,\ \beta_2 > 0$.

Proof of Lemma 7

We will prove Lemma 7 in the case $p = 1$. The case $p > 1$ works analogously for the representation (C-1). Recall from Assumption 2-*(ii)* that $\mathcal{F}_{t-1}^N = \sigma(\xi_{is} : 1 \leq i \leq N, s \leq t-1)$. Then, for $N \in \mathbb{N}$, we have that $\mathrm{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \mathrm{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}^N)$, see for example Shiryaev (2016, p. 210). Before proving each single point of the Lemma 7 we proof the following helpful results.

Lemma 9. Rewrite the linear model (4.2) as $\mathbf{Y}_t = f(\mathbf{Y}_{t-1}, \boldsymbol{\theta}) + \boldsymbol{\xi}_t$, for $t \geq 0$ where $\boldsymbol{\xi}_t = \mathbf{Y}_t - \boldsymbol{\lambda}_t$ and $f(\mathbf{Y}_{t-1}, \boldsymbol{\theta}) = \boldsymbol{\lambda}_t = \boldsymbol{\beta}_0 + \mathbf{G}\mathbf{Y}_{t-1}$. Define the following predictors, for $J > 0$:

$$\bar{\mathbf{Y}}_t = \begin{cases} f(\bar{\mathbf{Y}}_{t-1}, \boldsymbol{\theta}), & t > 0 \\ \mathbf{Y}_0, & t \leq 0 \end{cases}, \qquad \hat{\mathbf{Y}}_{t-J}^s = \begin{cases} f(\hat{\mathbf{Y}}_{t-J}^{s-1}, \boldsymbol{\theta}) + \boldsymbol{\xi}_s, & \max\{t-J, 0\} < s \leq t \\ \bar{\mathbf{Y}}_s, & s \leq \max\{t-J, 0\} \end{cases},$$

where $f(\bar{\mathbf{Y}}_{t-1}, \boldsymbol{\theta}) = \boldsymbol{\beta}_0 + \mathbf{G}\bar{\mathbf{Y}}_{t-1}$ and $f(\hat{\mathbf{Y}}_{t-J}^{t-1}, \boldsymbol{\theta}) = \hat{\boldsymbol{\lambda}}_{t-J}^t = \boldsymbol{\beta}_0 + \mathbf{G}\hat{\mathbf{Y}}_{t-J}^{t-1}$. Let $\tilde{\mathbf{Y}}_t^* = c\mathbf{Y}_t + (1-c)\bar{\mathbf{Y}}_t$ and $\tilde{\mathbf{Y}}_t = c\mathbf{Y}_t + (1-c)\hat{\mathbf{Y}}_{t-J}^t$ with $0 \leq c \leq 1$. Then,

$$\left|\mathbf{Y}_t - \hat{\mathbf{Y}}_{t-J}^t\right|_\infty \leq d^J \sum_{j=0}^{t-J-1} d^j \left|\boldsymbol{\xi}_{t-J-j}\right|_\infty ,$$

where $|\boldsymbol{\xi}_t|_\infty = \max_{1 \leq j \leq N} |\xi_{it}|$.

Proof. Set $t \geq 0$,

$$\begin{aligned}
\left|\mathbf{Y}_t - \bar{\mathbf{Y}}_t\right|_\infty &= \left|f(\mathbf{Y}_{t-1}, \boldsymbol{\theta}) + \boldsymbol{\xi}_t - f(\bar{\mathbf{Y}}_{t-1}, \boldsymbol{\theta})\right|_\infty \\
&\leq \left|\left|\left| \frac{\partial}{\partial \mathbf{Y}} f(\tilde{\mathbf{Y}}_{t-1}^*, \boldsymbol{\theta}) \right|\right|\right|_\infty |\mathbf{Y}_{t-1} - \bar{\mathbf{Y}}_{t-1}|_\infty + |\boldsymbol{\xi}_t|_\infty \\
&\leq d \left|\mathbf{Y}_{t-1} - \bar{\mathbf{Y}}_{t-1}\right|_\infty + |\boldsymbol{\xi}_t|_\infty \\
&\leq d^2 \left|\mathbf{Y}_{t-2} - \bar{\mathbf{Y}}_{t-2}\right|_\infty + d \left|\boldsymbol{\xi}_{t-1}\right|_\infty + |\boldsymbol{\xi}_t|_\infty \\
&\;\;\vdots \\
&\leq d^t \left|\mathbf{Y}_0 - \bar{\mathbf{Y}}_0\right|_\infty + \sum_{j=0}^{t-1} d^j \left|\boldsymbol{\xi}_{t-j}\right|_\infty \\
&= \sum_{j=0}^{t-1} d^j \left|\boldsymbol{\xi}_{t-j}\right|_\infty .
\end{aligned}$$

The first inequality holds for an application of the multivariate mean value theorem. Moreover, $\partial f(\mathbf{Y}_{t-1}, \boldsymbol{\theta})/\partial \mathbf{Y} = \mathbf{G}$ and $|||\mathbf{G}|||_\infty \leq \beta_1 + \beta_2 = d < 1$. Now set $t - J > 0$,

$$\begin{aligned}
\left|\mathbf{Y}_t - \hat{\mathbf{Y}}_{t-J}^t\right|_\infty &= \left|f(\mathbf{Y}_{t-1}, \boldsymbol{\theta}) + \boldsymbol{\xi}_t - f(\hat{\mathbf{Y}}_{t-J}^{t-1}, \boldsymbol{\theta}) - \boldsymbol{\xi}_t\right|_\infty \\
&\leq \left|\left|\left| \frac{\partial f(\tilde{\mathbf{Y}}_{t-1}, \boldsymbol{\theta})}{\partial \mathbf{Y}} \right|\right|\right|_\infty \left|\mathbf{Y}_{t-1} - \hat{\mathbf{Y}}_{t-J}^{t-1}\right|_\infty \\
&\leq d \left|\mathbf{Y}_{t-1} - \hat{\mathbf{Y}}_{t-J}^{t-1}\right|_\infty \\
&\leq d^2 \left|\mathbf{Y}_{t-2} - \hat{\mathbf{Y}}_{t-J}^{t-2}\right|_\infty \\
&\;\;\vdots \\
&\leq d^J \left|\mathbf{Y}_{t-J} - \bar{\mathbf{Y}}_{t-J}\right|_\infty \\
&\leq d^J \sum_{j=0}^{t-J-1} d^j \left|\boldsymbol{\xi}_{t-J-j}\right|_\infty
\end{aligned}$$

and the last inequality comes from the previous recursion. It is immediate to see that, for $t-J<0$, $\left|\mathbf{Y}_t-\hat{\mathbf{Y}}_{t-J}^t\right|_\infty \leq d^{J-t}\left|\mathbf{Y}_0-\bar{\mathbf{Y}}_0\right|_\infty = 0$. □

Proof of (1)

Define $\mathbf{W}_t=(\mathbf{Y}_t,\mathbf{Y}_{t-1})^T$, $\hat{\mathbf{W}}_{t-J}^t=(\hat{\mathbf{Y}}_{t-J}^t,\hat{\mathbf{Y}}_{t-J}^{t-1})^T \coloneqq f(\boldsymbol{\xi}_t,\ldots,\boldsymbol{\xi}_{t-J})$, $\hat{Y}_{it}$, $\hat{\lambda}_{it}$ the i-th elements of $\hat{\mathbf{Y}}_{t-J}^t$ and $\hat{\boldsymbol{\lambda}}_{t-J}^t$. Consider the following triangular array $\{g_{Nt}(\mathbf{W}_t):1\leq t\leq T_N, N\geq 1\}$, where $T_N\to\infty$ as $N\to\infty$. For any $\eta\in\mathbb{R}^m$, $g_{Nt}(\mathbf{W}_t)=N^{-1}\eta^T\frac{\partial\boldsymbol{\lambda}_t^T}{\partial\boldsymbol{\theta}}\mathbf{C}_t\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}^T}\eta=\sum_{r=1}^m\sum_{l=1}^m\eta_r\eta_l h_{rlt}$ where $N^{-1}\mathbf{h}_{Nt}=(h_{rlt})_{1\leq r,l\leq m}$. We take the most complicated element, h_{22t}, the result is analogously proven for the other elements. Define $l_{1it}=\left|(w_i^T\mathbf{Y}_{t-1})^2Y_{it}(\hat{\lambda}_{it}+\lambda_{it})\right|$, $l_{2it}=\left|(w_i^T\mathbf{Y}_{t-1})^2\lambda_{it}^2\right|$ and $l_{3it}=\left|\hat{Y}_{it}\lambda_{it}^2(Y_{it-1}+\hat{Y}_{it-1})\sum_{j=1}^N w_{ij}(Y_{jt-1}+\hat{Y}_{jt-1})\right|$. Additionally, the equality $\left|\hat{\lambda}_{it}-\lambda_{it}\right|=\left|Y_{it}-\hat{Y}_{it}\right|$ is a consequence of the constructions in Lemma 9. Then

$$
\begin{aligned}
\left|h_{22t}-h_{22,t-J}^t\right| &= \left|\frac{1}{N}\sum_{i=1}^N\frac{(w_i^T\mathbf{Y}_{t-1})^2Y_{it}}{\lambda_{it}^2}-\frac{1}{N}\sum_{i=1}^N\frac{(w_i^T\hat{\mathbf{Y}}_{t-J}^{t-1})^2\hat{Y}_{it}}{\hat{\lambda}_{it}^2}\right| \\
&\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^N\left|(w_i^T\mathbf{Y}_{t-1})^2Y_{it}\hat{\lambda}_{it}^2-(w_i^T\hat{\mathbf{Y}}_{t-J}^{t-1})^2\hat{Y}_{it}\lambda_{it}^2\right| \\
&\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^N\left|(w_i^T\mathbf{Y}_{t-1})^2Y_{it}(\hat{\lambda}_{it}^2-\lambda_{it}^2)+\left[(w_i^T\mathbf{Y}_{t-1})^2Y_{it}-(w_i^T\hat{\mathbf{Y}}_{t-J}^{t-1})^2\hat{Y}_{it}\right]\lambda_{it}^2\right| \\
&\leq \frac{\beta_0^{-4}}{N}\left|\sum_{i=1}^N(w_i^T\mathbf{Y}_{t-1})^2Y_{it}(\hat{\lambda}_{it}+\lambda_{it})(\hat{\lambda}_{it}-\lambda_{it})\right|+\frac{\beta_0^{-4}}{N}\left|\sum_{i=1}^N(w_i^T\mathbf{Y}_{t-1})^2\lambda_{it}^2(Y_{it}-\hat{Y}_{it})\right| \\
&\quad+\frac{\beta_0^{-4}}{N}\left|\sum_{i=1}^N\hat{Y}_{it}\lambda_{it}^2\left[(w_i^T\mathbf{Y}_{t-1})^2-(w_i^T\hat{\mathbf{Y}}_{t-J}^{t-1})^2\right]\right| \\
&\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{1it}\left|\hat{\lambda}_{it}-\lambda_{it}\right|+\frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{2it}\left|Y_{it}-\hat{Y}_{it}\right| \\
&\quad+\frac{\beta_0^{-4}}{N}\sum_{i=1}^N\hat{Y}_{it}\lambda_{it}^2\left|(w_i^T\mathbf{Y}_{t-1})+(w_i^T\hat{\mathbf{Y}}_{t-J}^{t-1})\right|\left|(w_i^T\mathbf{Y}_{t-1})-(w_i^T\hat{\mathbf{Y}}_{t-J}^{t-1})\right| \\
&\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{1it}\left|\hat{\lambda}_{it}-\lambda_{it}\right|+\frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{2it}\left|Y_{it}-\hat{Y}_{it}\right| \\
&\quad+\frac{\beta_0^{-4}}{N}\sum_{i=1}^N\hat{Y}_{it}\lambda_{it}^2\left|\sum_{j=1}^N w_{ij}(Y_{jt-1}+\hat{Y}_{jt-1})\right|\left|\sum_{j=1}^N w_{ij}(Y_{jt-1}-\hat{Y}_{jt-1})\right| \\
&\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^N(l_{1it}+l_{2it})\left|Y_{it}-\hat{Y}_{it}\right|+\frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{3it}\left|\sum_{j=1}^N w_{ij}(Y_{jt-1}-\hat{Y}_{jt-1})\right|.
\end{aligned}
$$

Set $1/a+1/b=1/2$ and $1/q+1/p+1/n=1/a$. By Cauchy-Schwarz inequality, as $w_{ij}>0$ for $j=1,\ldots,N$ and $\sum_{j=1}^N w_{ij}=1$ we have that $(w_i^T\mathbf{Y}_{t-1})^2=\left(\sum_{j=1}^N w_{ij}Y_{jt-1}\right)^2\leq\sum_{j=1}^N w_{ij}Y_{jt-1}^2$. So, $\max_{1\leq i\leq N}\left\|(w_i^T\mathbf{Y}_{t-1})^2\right\|_q\leq\max_{1\leq i\leq N}\left(\sum_{j=1}^N w_{ij}\left\|Y_{jt-1}^2\right\|_q\right)\leq\max_{i\geq 1}\left\|Y_{it}^2\right\|_q\leq C_{2q}^{1/q}<\infty$, by Proposition 7. Moreover, $\max_{i\geq1}\left\|\lambda_{it}^2\right\|_n\leq\max_{i\geq1}\left\|Y_{it}^2\right\|_n\leq C_n$, by the conditional Jensen's inequality. Similarly, $\max_{i\geq1}\left\|\hat{\lambda}_{it}^2\right\|_n\leq\max_{i\geq1}\left\|\hat{Y}_{it}^2\right\|_n$. An application of Lemma 9 provides $\max_{i\geq1}\left\|Y_{it}-\hat{Y}_{it}\right\|_b\leq d^J\sum_{j=0}^{t-J-1}d^j\max_{i\geq1}\|\xi_{it}\|_b\leq d^J 2C_b^{1/b}/(1-d)$. By an analogous recursion argument, it holds that $\max_{i\geq1}\left\|\hat{Y}_{it}^2\right\|_n\leq 2\beta_0\sum_{j=0}^\infty d^j+\sum_{j=0}^\infty d^j\max_{i\geq1}\|\xi_{it}\|_n\leq(2\beta_0+2C_n^{1/n})/(1-d)\coloneqq$

$\Delta < \infty$. By Holder's inequality $\max_{i\geq 1} \|l_{1it}\|_a \leq \max_{i\geq 1} \left\|(w_i^T \mathbf{Y}_{t-1})^2\right\|_q \|Y_{it}\|_p \left(\left\|\hat{\lambda}_{it}\right\|_n + \left\|\hat{\lambda}_{it}\right\|_n\right) < l_1 < \infty$. In the same way we can conclude that $\max_{i\geq 1} \|l_{2it}\|_q < l_2 < \infty$ and $\max_{i\geq 1} \|l_{3it}\|_q < l_3 < \infty$. Then, by Minkowski inequality

$$\begin{aligned}
\left\|h_{22t} - h^t_{22,t-J}\right\|_2 &\leq \frac{\beta_0^{-4}}{N} \sum_{i=1}^N \|l_{1it} + l_{2it}\|_a \left\|Y_{it} - \hat{Y}_{it}\right\|_b + \frac{\beta_0^{-4}}{N} \sum_{i=1}^N \|l_{3it}\|_a \sum_{j=1}^N w_{ij} \left\|Y_{jt-1} - \hat{Y}_{jt-1}\right\|_b \\
&\leq \beta_0^{-4} \max_{1\leq i\leq N} \left(\|l_{1it}\|_a + \|l_{2it}\|_a\right) \left\|Y_{it} - \hat{Y}_{it}\right\|_b + \beta_0^{-4} \max_{1\leq i\leq N} \|l_{3it}\|_a \left\|Y_{it-1} - \hat{Y}_{it-1}\right\|_b \\
&\leq \beta_0^{-4} \left(l_1 + l_2 + l_3\right) 2C_b^{1/b} d^{J-1} \sum_{j=0}^{t-J-1} d^j \leq \frac{\beta_0^{-4} \left(l_1 + l_2 + l_3\right) 2C_b^{1/b}}{1-d} d^{J-1} := c_{22}\nu_J \,,
\end{aligned}$$

with $\nu_J = d^{J-1}$. By the definition in Assumption 2-*(ii)*, recall $\mathcal{F}^N_{t-J,t+J} = \sigma\left(\xi_{it} : 1 \leq i \leq N, t-J \leq t \leq t+J\right)$. Since $\mathrm{E}\left[g_{Nt}(\mathbf{W}_t)|\mathcal{F}^N_{t-J,t+J}\right]$ is the optimal $\mathcal{F}^N_{t-J,t+J}$-measurable approximation to $g_{Nt}(\mathbf{W}_t)$ in the L^2-norm and $g_{Nt}(\hat{\mathbf{W}}^t_{t-J})$ is $\mathcal{F}^N_{t-J,t+J}$-measurable, it follows that

$$\begin{aligned}
\left\|g_{Nt}(\mathbf{W}_t) - \mathrm{E}\left[g_{Nt}(\mathbf{W}_t)|\mathcal{F}^N_{t-J,t+J}\right]\right\|_2 &\leq \left\|g_{Nt}(\mathbf{W}_t) - g_{Nt}(\hat{\mathbf{W}}^t_{t-J})\right\|_2 \\
&\leq \sum_{r=1}^m \sum_{l=1}^m \eta_k \eta_l \left\|h_{rlt} - \hat{h}^t_{rl,t-J}\right\|_2 \\
&\leq c_{Nt}\nu_j \,,
\end{aligned}$$

where $c_{Nt} = \sum_{r=1}^m \sum_{l=1}^m \eta_r \eta_l c_{rl}$ and $\nu_J = d^{J-1} \to 0$ as $J \to \infty$, establishing L^p-near epoch dependence (L^p-NED), with $p \in [1,2]$, for the triangular array $\{X_{Nt} = g_{Nt}(\mathbf{W}_t) - \mathrm{E}\left[g_{Nt}(\mathbf{W}_t)\right]\}$; see Andrews (1988). Moreover, for a similar argument above, it is easy to see that $\mathrm{E}\left|X_{Nt}\right|^2 < \infty$. Then, for Assumption 2-*(ii)* and the argument in Andrews (1988, p. 464), we have that $\{X_{Nt}\}$ is a uniformly integrable L^1-mixingale. Furthermore, since $\lim_{N\to\infty} T_N^{-1} \sum_{t=1}^{T_N} c_{Nt} < \infty$ the law of large number of Theorem 2 in Andrews (1988) provides the desired result $(NT)^{-1}\eta^T \mathbf{H}_{NT} \eta \xrightarrow{p} \eta^T \mathbf{H} \eta$ as $\min\{N,T\} \to \infty$. □

Proof of (2)

Let $\tilde{g}_{Nt}(\mathbf{W}_t) = N^{-1}\eta^T \frac{\partial \boldsymbol{\lambda}_t^T}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1} \boldsymbol{\Sigma}_t \mathbf{D}_t^{-1} \frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T} \eta = \sum_{r=1}^m \sum_{l=1}^m \eta_r \eta_l b_{rlt}$ where $N^{-1}\mathbf{b}_{Nt} = (b_{rlt})_{1\leq r,l\leq m}$ and $\boldsymbol{\Sigma}_t = \mathrm{E}(\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T | \mathcal{F}_{Nt-1})$, with $\boldsymbol{\xi}_t = \mathbf{Y}_t - \boldsymbol{\lambda}_t = \hat{\mathbf{Y}}^t_{t-J} - \hat{\boldsymbol{\lambda}}^t_{t-J}$, since $\mathrm{E}(\hat{\mathbf{Y}}^t_{t-J}|\mathcal{F}_{Nt-1}) = \hat{\boldsymbol{\lambda}}^t_{t-J}$. We consider again the most complicated element, that is b_{22t}. For $1 \leq i,j \leq N$, define $\sigma_{ijt} = \mathrm{E}(\xi_{it}\xi_{jt}|\mathcal{F}_{Nt-1})$, then

$$\begin{aligned}
\left|b_{22t} - b^t_{22,t-J}\right| &= \left| \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \frac{(w_i^T \mathbf{Y}_{t-1})(w_j^T \mathbf{Y}_{t-1})\sigma_{ijt}}{\lambda_{it}\lambda_{jt}} - \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \frac{(w_i^T \hat{\mathbf{Y}}^{t-1}_{t-J})(w_j^T \hat{\mathbf{Y}}^{t-1}_{t-J})\sigma_{ijt}}{\hat{\lambda}_{it}\hat{\lambda}_{jt}} \right| \\
&\leq \beta_0^{-4} \left| \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \sigma_{ijt} \left| (w_i^T \mathbf{Y}_{t-1})(w_j^T \mathbf{Y}_{t-1})\hat{\lambda}_{it}\hat{\lambda}_{jt} - (w_i^T \hat{\mathbf{Y}}^{t-1}_{t-J})(w_j^T \hat{\mathbf{Y}}^{t-1}_{t-J})\lambda_{it}\lambda_{jt} \right| \right| \\
&\leq \beta_0^{-4} \left| \frac{1}{N} \sum_{i,j=1}^N \sigma_{ijt} \left(r_{1ijt} \left|\lambda_{it} - \hat{\lambda}_{it}\right| + r_{2ijt} \left| \sum_{j=1}^N w_{ij} (Y_{jt-1} - \hat{Y}_{jt-1}) \right| \right) \right| .
\end{aligned}$$

The second inequality is obtained as for the element of the Hessian h_{22t} in the previous section. Moreover, $r_{1ijt} = (w_i^T \mathbf{Y}_{t-1})(w_j^T \mathbf{Y}_{t-1})(\hat{\lambda}_{jt} + \lambda_{jt})$ and $r_{2ijt} = \lambda_{it}\lambda_{jt}(w_i^T \mathbf{Y}_{t-1} + w_i^T \mathbf{Y}^{t-1}_{t-J})$ and $\frac{1}{N}\sum_{i,j=1}^N \|\sigma_{ijt}\|_a \leq \frac{1}{N}\sum_{i,j=1}^N \|\xi_{it}\xi_{jt}\|_a < \lambda < \infty$ for Assumption 2-*(i)*. Set $1/q + 1/h = 1/b$. Note that $\max_{i,j\geq 1} \|r_{1ijt}\|_q < r_1 < \infty$, $\max_{i,j\geq 1} \|r_{2ijt}\|_q < r_2 < \infty$

by the same argument of $\max_{i\geq 1}\|l_{1it}\|_a < l_1$ above. Then,

$$\begin{aligned}\left\|b_{22t}-b^t_{22,t-J}\right\|_2 &\leq \beta_0^{-4}\frac{1}{N}\sum_{i,j=1}^{N}\|\sigma_{ijt}\|_a\left\|r_{1ijt}\left|\lambda_{it}-\hat{\lambda}_{it}\right|+r_{2ijt}\left|\sum_{j=1}^{N}w_{ij}(Y_{jt-1}-\hat{Y}_{jt-1})\right|\right\|_b\\ &\leq \beta_0^{-4}\lambda\max_{1\leq i,j\leq N}\|r_{1ijt}\|_q\left\|Y_{it}-\hat{Y}_{it}\right\|_h+\beta_0^{-4}\lambda\max_{1\leq i,j\leq N}\|r_{2ijt}\|_q\left\|Y_{it-1}-\hat{Y}_{it-1}\right\|_h\\ &\leq \frac{\beta_0^{-4}\lambda\,(r_1+r_2)\,2C_h^{1/h}}{1-d}d^{J-1}\coloneqq r_{22}\nu_J\,.\end{aligned}$$

Here again $\nu_J = d^{J-1}$. Then, the triangular array $\left\{\tilde{X}_{Nt}=\tilde{g}_{Nt}(\mathbf{W}_t)-\mathrm{E}\left[\tilde{g}_{Nt}(\mathbf{W}_t)\right]\right\}$ is L^p-NED and Theorem 2 in Andrews (1988) holds for it. This result and Assumption 1-*(ii)* yields to the convergence

$$(NT)^{-1}\eta^T\mathbf{B}_{NT}\eta\xrightarrow{p}\eta^T\mathbf{B}\eta\,, \tag{C-3}$$

as $\min\{N,T\}\to\infty$, for any $\eta\in\mathbb{R}^m$.
Now we show asymptotic normality. Define $\varepsilon_{Nt}=\eta^T\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}}^T\mathbf{D}_t^{-1}\boldsymbol{\xi}_t$, and recall the σ-field $\mathcal{F}_t^N=\sigma\left(\xi_{is}:1\leq i\leq N, s\leq t\right)$. Set $S_{Nt}=\sum_{s=1}^{t}\varepsilon_{Ns}$, so $\left\{S_{Nt},\mathcal{F}_t^N:t\leq T_N, N\geq 1\right\}$ is a martingale array. Following a similar argument above, for Cauchy-Schwarz inequality and Assumption 2-*(i)*, $\mathrm{E}\left(N^{-1}\eta^T\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}}^T\mathbf{D}_t^{-1}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^T\mathbf{D}_t^{-1}\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}^T}\eta\right)^2<C\beta_0^{-4}\lambda^2<\infty$, where $C=C_h\sum_{r,l=1}^{m}|\eta_r|^2|\eta_l|^2$, satisfying the Lindberg's condition

$$\frac{1}{NT_N}\sum_{t=1}^{T_N}\mathrm{E}\left[\varepsilon_{Nt}^2 I\left(|\varepsilon_{Nt}|>\sqrt{NT_N}\delta\right)\mid\mathcal{F}_{t-1}^N\right]\leq\frac{\delta^{-2}}{N^2T_N^2}\sum_{t=1}^{T_N}\mathrm{E}\left(\varepsilon_{Nt}^4\mid\mathcal{F}_{t-1}^N\right)\xrightarrow{p}0\,,$$

for any $\delta>0$, as $N\to\infty$. By the result in equation (C-3)

$$\frac{1}{NT_N}\sum_{t=1}^{T_N}\mathrm{E}\left(\varepsilon_{Nt}^2\mid\mathcal{F}_{t-1}^N\right)=\frac{1}{NT_N}\sum_{t=1}^{T_N}\eta^T\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}}^T\mathbf{D}_t^{-1}\mathrm{E}(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^T\mid\mathcal{F}_{t-1}^N)\mathbf{D}_t^{-1}\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}^T}\eta\xrightarrow{p}\eta^T\mathbf{B}\eta\,,$$

for any $\delta>0$, as $N\to\infty$. Then, the central limit theorem for martingale array in Hall and Heyde (1980, Cor. 3.1) applies, $(NT_N)^{-1/2}S_{NT_N}\xrightarrow{d}N(0,\eta^T\mathbf{B}\eta)$, leading to the desired result. □

Proof of (3)

Consider the third derivative

$$\frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_l\partial\boldsymbol{\theta}_k}=2\frac{Y_{i,t}}{\lambda_{i,t}^3(\boldsymbol{\theta})}\left(\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_l}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_k}\right)\,.$$

Take the case $\boldsymbol{\theta}_j=\boldsymbol{\theta}_l=\boldsymbol{\theta}_k=\beta_1$, the proof is analogous for the other derivatives,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial\beta_1^3}=\frac{1}{N}\sum_{i=1}^{N}2\frac{Y_{i,t}}{\lambda_{i,t}^3(\boldsymbol{\theta})}\left(w_i^T\mathbf{Y}_{t-1}\right)^3\leq\frac{1}{N}\sum_{i=1}^{N}2\beta_0^{-3}Y_{i,t}\left(w_i^T\mathbf{Y}_{t-1}\right)^2\left(w_i^T\mathbf{Y}_{t-1}\right)\coloneqq\frac{1}{N}\sum_{i=1}^{N}m_{i,t}\,.$$

Now, define $M_{NT}\coloneqq\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}m_{i,t}$ and $\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}(m_{i,t})<\infty$ since all the moment of $\mathbf{Y}_t$ exist. It is easy to see that $M_{NT}\xrightarrow{p}M$ as $\min\{N,T\}\to\infty$, similarly as above for point (1) and (2), then point (3) of Lemma 7 follows by Assumption 1-*(iii)*. We omit the details. □

Proof of Lemma 8

The proof is analogous to that of Lemma 7. We will point out only the parts which differ significantly.

Lemma 10. Define $\mathbf{Z}_t = \log(1 + \mathbf{Y}_t)$. Rewrite the linear model (4.6) as $\mathbf{Z}_t = \boldsymbol{\nu}_t + \boldsymbol{\psi}_t$, for $t \geq 0$, where $\boldsymbol{\nu}_t = \boldsymbol{\beta}_0 + \mathbf{G}\mathbf{Z}_{t-1}$. Define the predictors $\hat{\mathbf{Z}}^t_{t-J} = \hat{\boldsymbol{\nu}}^t_{t-J} + \boldsymbol{\psi}_t$, where $\hat{\boldsymbol{\nu}}^t_{t-J} = \boldsymbol{\beta}_0 + \mathbf{G}\hat{\mathbf{Z}}^{t-1}_{t-J}$ and $\bar{\mathbf{Z}}^t_{t-J}$ analogously to Lemma 9. Then, $\left|\mathbf{Z}_t - \hat{\mathbf{Z}}^t_{t-J}\right|_\infty \leq d^J \sum_{j=0}^{t-J-1} d^j \left|\boldsymbol{\psi}_{t-J-j}\right|_\infty$.

Proof. The proof is analogous to Lemma 9 and therefore is omitted. □

Proof of (1)

Set $\hat{\mathbf{Y}}^t_{t-J} = \exp(\hat{\boldsymbol{\nu}}^t_{t-J}) + \boldsymbol{\xi}_t$, $\mathbf{W}_t = (\mathbf{Z}_t, \mathbf{Z}_{t-1}, \mathbf{Y}_t)^T$, $\hat{\mathbf{W}}^t_{t-J} = (\hat{\mathbf{Z}}^t_{t-J}, \hat{\mathbf{Z}}^{t-1}_{t-J}, \hat{\mathbf{Y}}^t_{t-J})^T \coloneqq f(\boldsymbol{\psi}_t, \ldots, \boldsymbol{\psi}_{t-J})$. Consider the triangular array $\{g_{Nt}(\mathbf{W}_t) : 1 \leq t \leq T_N; N \geq 1\}$, where $T_N \to \infty$ as $N \to \infty$. For any $\eta \in \mathbb{R}^m$, $g_{Nt}(\mathbf{W}_t) = N^{-1}\eta^T \frac{\partial \boldsymbol{\nu}_t^T}{\partial \boldsymbol{\theta}} \mathbf{D}_t \frac{\partial \boldsymbol{\nu}_t}{\partial \boldsymbol{\theta}^T}\eta = \sum_{r=1}^m \sum_{l=1}^m \eta_r \eta_l h_{rlt}$. Then,

$$\begin{aligned}
\left|h_{22t} - h^t_{22,t-J}\right| &= \left|\frac{1}{N}\sum_{i=1}^N (w_i^T \mathbf{Z}_{t-1})^2 \exp(\nu_{it}) - \frac{1}{N}\sum_{i=1}^N (w_i^T \hat{\mathbf{Z}}^{t-1}_{t-J})^2 \exp(\hat{\nu}_{it})\right| \\
&\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^N c^*_{1it} \left|\exp(\nu_{it}) - \exp(\hat{\nu}_{it})\right| + \frac{\beta_0^{-4}}{N}\sum_{i=1}^N c_{2it}\left|\sum_{j=1}^N w_{ij}(Z_{jt-1} - \hat{Z}_{jt-1})\right| \\
&\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^N c^*_{1it} \exp(\nu_{it})\exp(\hat{\nu}_{it}) \left|\nu_{it} - \hat{\nu}_{it}\right| + \frac{\beta_0^{-4}}{N}\sum_{i=1}^N c_{2it}\left|\sum_{j=1}^N w_{ij}(Z_{jt-1} - \hat{Z}_{jt-1})\right|,
\end{aligned}$$

where $c^*_{1it} = (w_i^T \mathbf{Z}_{t-1})^2$, $c_{1it} = c^*_{1it}\exp(\nu_{it})\exp(\hat{\nu}_{it})$ and $c_{2it} = \exp(\hat{\nu}_{it})(w_i^T \mathbf{Z}_{t-1} + w_i^T \hat{\mathbf{Z}}^{t-1}_{t-J})$. The second inequality follows by $|\exp(x) - \exp(y)| = |\exp(y)(\exp(x-y) - 1)|$ and $|(\exp(x-y) - 1)| \leq |\exp(x-y)|\,|x-y| \leq |\exp(x)|\,|x-y|$, for $x, y \in \mathbb{R}^+_0$. Set Set $1/a + 1/b = 1/2$ and $1/q + 1/p + 1/n = 1/a$. It is easy to show that $\max_{1\leq i\leq N} \left\|(w_i^T \mathbf{Z}_{t-1})^2\right\|_q \leq \max_{1\leq i\leq N} \left\|Z^2_{it}\right\|_q$, by Cauchy-Schwarz inequality. Moreover, $\max_{i\geq 1} \|Z_{it}\|_q \leq \max_{i\geq 1} \|Y_{it}\|_q$ and $\max_{i\geq 1} \|\nu_{it}\|_q \leq \beta_0 + (\beta_1 + \beta_2)\max_{i\geq 1} \|Z_{it}\|_q$. All these quantities are bounded by Assumption 2′-*(i)*. Lemma 10 implies $\max_{i\geq 1} \left\|Z_{it} - \hat{Z}_{it}\right\|_b = \max_{i\geq 1} \|\nu_{it} - \hat{\nu}_{it}\|_b \leq d^J \sum_{j=0}^{t-J-1} d^j \max_{i\geq 1} \|\psi_{it}\|_b \leq d^J C$, where C is some constant, and $\max_{i\geq 1} \left\|\hat{Z}_{it}\right\|_q \leq 2\beta_0 \sum_{j=0}^\infty d^j + \sum_{j=0}^\infty d^j \max_{i\geq 1} \|\psi_{it}\|_q < \Delta < \infty$, again by Assumption 2′-*(i)*. Define e_i a vector of zero's with 1 only in the i-th position. Moreover, recall that $\hat{\nu}_{it} = \sum_{j=0}^{t-1} e_i^T \mathbf{G}^j \boldsymbol{\beta}_0 + \sum_{j=1}^{J-1} e_i^T \mathbf{G}^j \boldsymbol{\psi}_{t-j} = b_0 + \sum_{j=1}^{J-1} f(\boldsymbol{\theta}, w_{ij})\psi_{jt-j}$, where $b_0 = \sum_{j=0}^{t-1} e_i^T \mathbf{G}^j \boldsymbol{\beta}_0$ and $f(\boldsymbol{\theta}, w_{ij}) \in \mathbb{R}$ is some deterministic continuous function. Then

$$\begin{aligned}
\mathrm{E}\left[\exp(q\hat{\nu}_{it})\right] &\leq \exp(b_0 q)\mathrm{E}\left[\exp\left(q\sum_{j=1}^{J-1} f(\boldsymbol{\theta}, w_{ij})\psi_{jt-j}\right)\right] \\
&\leq \exp(b_0 q)\prod_{j=1}^{J-1} \mathrm{E}^{1/2l_j}\left[\exp\left(c_j\left|\psi_{jt-j}\right|\right)\right],
\end{aligned}$$

where the second inequality works for successive use of Cauchy-Schwarz inequality, with $1 \leq l_{J-1} \leq J-3$ and $c_j = 2l_j\left|f(\boldsymbol{\theta}, w_{ij})\right| q$. Now note that $\mathrm{E}[\exp(c_j\left|\psi_{jt}\right|)] \leq \mathrm{E}^{1/2}[\exp(2c_j\left|Z_{jt}\right|)]\mathrm{E}^{1/2}[\exp(2c_j\left|\nu_{jt}\right|)]$, and, by an application of the binomial theorem, we show that

$$\mathrm{E}\left[\exp\left(\left|Z_{jt}\right|\right)^{2c_j}\right] = \mathrm{E}\left[\left(1 + Y_{jt}\right)^{2c_j}\right] \leq \sum_{k=0}^{2c_j}\binom{2cj}{k}\mathrm{E}\left|\mathbf{Y}_{jt}\right|^k$$

is finite for Assumption 2′-*(i)*, as well as $\mathrm{E}[\exp(2c_j\left|\nu_{jt}\right|)] < \infty$, so $\left\|\exp(\hat{\nu}_{it})\right\|_q < \Delta_e < \infty$ and $\max_{i\geq 1}\left\|c_{1it}\right\|_q < c_1 < \infty$ and $\max_{i\geq 1}\left\|c_{2it}\right\|_q < c_2 < \infty$. Then, $\left\|h_{22t} - h^t_{22,t-J}\right\|_2 \leq c_{22}\nu_J$ is L^p-NED and, by Assumption 2′-*(ii)*, the conclusion follows as for the linear model. □

Proof of (2)

Let $\tilde{g}_{Nt}(\mathbf{W}_t) = N^{-1}\eta^T \frac{\partial \boldsymbol{\nu}_t^T}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}_t \frac{\partial \boldsymbol{\nu}_t}{\partial \boldsymbol{\theta}^T}\eta = \sum_{r=1}^m \sum_{l=1}^m \eta_r \eta_l b_{rlt}$, where $\boldsymbol{\Sigma}_t = \mathrm{E}(\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T | \mathcal{F}_{Nt-1})$, with $\boldsymbol{\xi}_t = \mathbf{Y}_t - \exp(\boldsymbol{\nu}_t) = \hat{\mathbf{Y}}_{t-J}^t - \exp(\hat{\boldsymbol{\nu}}_{t-J}^t)$, since $\mathrm{E}(\hat{\mathbf{Y}}_{t-J}^t | \mathcal{F}_{Nt-1}) = \exp(\hat{\boldsymbol{\nu}}_{t-J}^t)$. Analogously as above

$$\left| b_{22t} - b_{22,t-J}^t \right| \leq \left| \frac{1}{N} \sum_{i,j=1}^N \sigma_{ijt} \left| (n_{1it} + n_{2it}) \sum_{j=1}^N w_{ij} (Y_{jt-1} - \hat{Y}_{jt-1}) \right| \right| ,$$

where $n_{1it} + n_{2it} = w_i^T \mathbf{Z}_{t-1} + w_i^T \hat{\mathbf{Z}}_{t-J}^{t-1}$, $\max_{i \geq 1} \left\| n_{1it} + n_{2it} \right\|_q < \Delta < \infty$ and $\frac{1}{N} \sum_{i,j=1}^N \left\| \sigma_{ijt} \right\|_a < \lambda < \infty$, for Assumption $2'$, proving L^p-NED. The proof of asymptotic normality follows the same fashion of the linear model and therefore is omitted. □

Proof of (3)

Consider the third derivative

$$\frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_k} = 2Y_{i,t} \left(\frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} \frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_l} \frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} \right) .$$

Take the case $\boldsymbol{\theta}_j = \boldsymbol{\theta}_l = \boldsymbol{\theta}_k = \beta_1$,

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial \beta_1^3} = \frac{1}{N} \sum_{i=1}^N 2Y_{i,t} \left(w_i^T \mathbf{Z}_{t-1} \right)^3 \coloneqq \frac{1}{N} \sum_{i=1}^N m_{i,t} .$$

The rest of the proof is omitted as it is in the same style of part (1) and (2). □

Proof of Theorem 17

Consider the following inequality, for the single (k,l) element of the Hessian matrix.

$$\left| \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N \left(\frac{\partial^2 l_{i,t}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_l} - \frac{\partial^2 l_{i,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_l} \right) \right| \leq \left| \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N \frac{\partial^3 l_{i,t}(\boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_s} \right| \left| \hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_{0,s} \right| \leq M_{NT} \left| \hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_{0,s} \right| ,$$

which converges to 0, in probability, as $\min\{N,T\} \to \infty$. The second inequality holds for condition 3 in Lemma 7, $\boldsymbol{\theta}_s$ is the s-element of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\star$ is an intermediate point between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. This result, together with condition 1 in Lemma 7, provides $(NT)^{-1}\mathbf{H}_{NT}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{H}(\boldsymbol{\theta}_0)$. It is immediate to show that the result $(NT)^{-1}\hat{\mathbf{B}}_{NT}(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{B}(\boldsymbol{\theta}_0)$ holds, as $\min\{N,T\} \to \infty$. The proof is closely analogous to the proof of Lemma 7-(2), by substituting $\xi_{it}\xi_{jt}$ to σ_{ijt}. Then, we only need to verify that $\left| \left| \left| (NT)^{-1}(\hat{\mathbf{B}}_{NT}(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{B}}_{NT}(\boldsymbol{\theta}_0)) \right| \right| \right| \xrightarrow{p} 0$, where $|||\cdot|||$ is a suitable matrix norm. Consider the following inequalities, for the single (k,l) element of the matrices $\hat{\mathbf{B}}_{NT}(\cdot)$.

$$\left| \frac{1}{NT} \sum_{t=1}^T \sum_{i,j=1}^N \left(\frac{\partial l_{i,t}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k} \frac{\partial l_{j,t}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_l} - \frac{\partial l_{i,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_k} \frac{\partial l_{j,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \right) \right| \leq D_1 + D_2 ,$$

defined as

$$D_1 = \left| \frac{1}{NT} \sum_{t=1}^T \sum_{i,j=1}^N \frac{\partial l_{i,t}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k} \left(\frac{\partial l_{j,t}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_l} - \frac{\partial l_{j,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \right) \right| = \left| \frac{1}{NT} \sum_{t=1}^T \sum_{i,j=1}^N \frac{\partial l_{i,t}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k} \right| o_p(1) = \frac{S_{NT}^{(l)}(\hat{\boldsymbol{\theta}})}{T} o_p(1) = 0$$

where the second equality works for the continuous mapping theorem and the fourth equality is true since $S_{NT}^{(l)}(\hat{\boldsymbol{\theta}})$ is the l-element of $\mathbf{S}_{NT}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, and

$$D_2 = \left| \frac{1}{NT} \sum_{t=1}^{T} \sum_{i,j=1}^{N} \frac{\partial l_{j,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \left(\frac{\partial l_{i,t}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k} - \frac{\partial l_{i,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_k} \right) \right| = \left| \frac{1}{NT} \sum_{t=1}^{T} \sum_{i,j=1}^{N} \frac{\partial l_{j,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \frac{\partial^2 l_{i,t}(\boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_s} \left(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_{0,s} \right) \right|$$
$$= \left| \frac{1}{NT} \sum_{t=1}^{T} \frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \frac{\partial^2 l_t(\boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_s} \right| o_p(1) = \mathcal{O}_p(1) o_p(1) \,.$$

The second equality works for the mean-value theorem. The last equality is true if the following sufficient condition is satisfied (Van der Vaart, 2000, Ex. 2.6).

$$\mathrm{E} \left| \frac{1}{NT} \sum_{t=1}^{T} \frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \frac{\partial^2 l_t(\boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_s} \right| \leq \frac{1}{N} \mathrm{E} \left| \frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \frac{\partial^2 l_t(\boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_s} \right| \leq \frac{1}{N} \left\| \frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \right\|_2 \left\| \frac{\partial^2 l_t(\boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_s} \right\|_2 = \mathcal{O}(1) \,. \tag{C-4}$$

We show (C-4) for the most complicated case, when $\boldsymbol{\theta}_l = \beta_1$, the proof is the same for the other derivatives. Then,

$$\frac{1}{N} \mathrm{E} \left(\frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_l} \right)^2 = \frac{1}{N} \mathrm{E} \left[\sum_{i=1}^{N} \left(\frac{Y_{it}}{\lambda_{it}} - 1 \right) \frac{\partial \lambda_{it}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right]^2 = \frac{1}{N} \mathrm{E} \left[\sum_{i=1}^{N} \left(\frac{\xi_{it}}{\lambda_{it}} \right) w_i^T \mathbf{Y}_{t-1} \right]^2$$
$$= \frac{1}{N} \sum_{i,j=1}^{N} \mathrm{E} \left(\frac{\xi_{it} \xi_{jt}}{\lambda_{it} \lambda_{jt}} \tilde{Y}_{i,t-1} \tilde{Y}_{j,t-1} \right) = \frac{1}{N} \mathbf{B}_N^{(22)} = \mathcal{O}(1) \,,$$

where the last equality comes from Assumption 2-(ii). The second term for $\boldsymbol{\theta}_k = \boldsymbol{\theta}_s = \beta_1$ is

$$\frac{1}{N} \mathrm{E} \left(\frac{\partial^2 l_t(\boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_s} \right)^2 = \frac{1}{N} \mathrm{E} \left(\sum_{i=1}^{N} \frac{Y_{it}}{\lambda_{it}^2(\boldsymbol{\theta}^\star)} \tilde{Y}_{i,t-1}^2 \right)^2 \leq \frac{(\beta_0^\star)^{-4}}{N} \sum_{i,j=1}^{N} \mathrm{E} \left(Y_{it} Y_{jt} \sum_{h=1}^{N} \sum_{m=1}^{N} w_{ih} w_{jm} Y_{h,t-1}^2 Y_{m,t-1}^2 \right)$$
$$\leq \frac{(\beta_0^\star)^{-4}}{N} \sum_{i,j=1}^{N} \sum_{h=1}^{N} \sum_{m=1}^{N} w_{ih} w_{jm} \left\| Y_{it} Y_{jt} \right\|_2 \left\| Y_{h,t-1}^2 Y_{m,t-1}^2 \right\|_2$$
$$\leq \frac{(\beta_0^\star)^{-4}}{N} \sum_{i,j=1}^{N} \left\| Y_{it} Y_{jt} \right\|_2 \max_{h,m \geq 1} \left\| Y_{h,t-1}^2 \right\|_4 \left\| Y_{m,t-1}^2 \right\|_4 \sum_{h=1}^{N} \sum_{m=1}^{N} w_{ih} w_{jm}$$
$$\leq \frac{C}{N} \sum_{i,j=1}^{N} \left\| Y_{it} Y_{jt} \right\|_2 = \mathcal{O}(1)$$

where the first inequality works since $\lambda_{it}(\boldsymbol{\theta}^\star) \geq \beta_0^\star$ and for Cauchy-Schwarz inequality. The last inequality holds for Proposition 7 and the fact that $\sum_{h=1}^{N} w_{ih} = 1$. Then, (C-4) holds true and $D_2 \xrightarrow{p} 0$, as $\min\{N, T\} \to \infty$, implying that $(NT)^{-1} \hat{\mathbf{B}}_{NT}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{B}(\boldsymbol{\theta}_0)$, and this ends the proof. □

Further simulations results

We present here further comments and results from the simulation study reported in Sec. 4.4.1. In the situation of independence ($\rho = 0$) the QMLE reduces to the standard MLE. When N is big and T is small we see that QMLE provides satisfactory results (Table C-2, C-5). However, this is not always the case. As it was mentioned in Sec. 4.4.1 when dependence is present, the quasi likelihood (4.10) is a rough approximation to the true likelihood. Intuitively, increasing N should confirm the asymptotic results of Thm. 15. However, at the same time, it could lead to a more complex structure of dependence among variables and then the quasi-likelihood might not approximate the true likelihood. In particular, when $N \to \infty$ and T is small, care must be taken in the interpretation of obtained

estimates. This fact is also confirmed by the Tables C-1 and C-4 who illustrate better results when there exists moderate dependence among the count variables. Finally, if both the temporal size T and the network size N are reasonably large, then Thm. 15 applies. The usage of the Clayton copula instead of the Gaussian (Tables C-3 and C-6) provide slightly better results but they are generally in agreement with previous observations.

Figure C-5 shows a QQ-plot of the standardized estimators for the log-linear model of order 1, with Gaussian copula ($\rho = 0.5$) and $N = 100$. When T is small then, we observed a deviation from normality, especially on the right tail of the distribution. When both dimensions are large, then the approximation is more satisfactory. Clearly, by reducing dependence among count variables, we can obtain better large-sample approximations but these results are not plotted due to space constraints.

Table C-1: Estimators obtained from $S = 1000$ simulations of model (4.2), for various values of N and T. Data are generated by using the Gaussian copula with $\rho = 0.2$ and $p = 1$. Model (4.2) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	*AIC*	*BIC*	*QIC*
20	100	0.201	0.395	0.495	0.199	0.386	0.490	0.015	0.007			
		(0.020)	(0.041)	(0.030)	(0.021)	(0.049)	(0.033)	(0.043)	(0.022)	93.1	99.6	93.8
		100	100	100	100	100	100	0.2	0.2			
	200	0.201	0.399	0.497	0.199	0.382	0.493	0.012	0.005			
		(0.014)	(0.030)	(0.021)	(0.015)	(0.035)	(0.023)	(0.031)	(0.015)	92.7	99.9	93.9
		100	100	100	100	100	100	0.2	0.2			
100	10	0.229	0.337	0.478	0.219	0.363	0.468	0.025	0.012			
		(0.063)	(0.051)	(0.051)	(0.063)	(0.063)	(0.058)	(0.061)	(0.038)	88.8	90.2	87.6
		58.0	97.1	99.8	35.1	84.0	98.0	0.1	0.1			
	20	0.216	0.384	0.485	0.211	0.376	0.479	0.014	0.007			
		(0.045)	(0.037)	(0.035)	(0.045)	(0.046)	(0.040)	(0.043)	(0.026)	89.7	93.9	93.4
		99.6	100	100	98.9	99.9	100	0.1	0.3			
	100	0.203	0.396	0.496	0.201	0.392	0.492	0.007	0.003			
		(0.020)	(0.016)	(0.016)	(0.020)	(0.020)	(0.018)	(0.019)	(0.011)	86.8	96.6	94.6
		100	100	100	100	100	100	0.2	0.1			
	200	0.201	0.398	0.498	0.200	0.395	0.495	0.005	0.002			
		(0.014)	(0.012)	(0.011)	(0.014)	(0.014)	(0.013)	(0.013)	(0.008)	85.6	96.9	93.8
		100	100	100	100	100	100	0.2	0.3			

Table C-2: Estimators obtained from $S = 1000$ simulations of model (4.2), for various values of N and T. Data are generated by using the Gaussian copula with $\rho = 0$ and $p = 1$. Model (4.2) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	*AIC*	*BIC*	*QIC*
20	100	0.201	0.399	0.496	0.198	0.389	0.490	0.015	0.008			
		(0.014)	(0.039)	(0.025)	(0.015)	(0.046)	(0.028)	(0.042)	(0.020)	94.8	99.8	94.4
		100	100	100	100	100	100	0.4	0.2			
	200	0.201	0.400	0.498	0.199	0.392	0.493	0.012	0.005			
		(0.010)	(0.027)	(0.018)	(0.010)	(0.032)	(0.020)	(0.029)	(0.014)	94.9	100	94.9
		100	100	100	100	100	100	0.4	0.0			
100	10	0.203	0.397	0.497	0.196	0.385	0.487	0.018	0.012			
		(0.027)	(0.037)	(0.031)	(0.030)	(0.046)	(0.036)	(0.047)	(0.030)	94.4	96.0	91.4
		99.8	100	100	99.2	100	100	0.3	0.2			
	20	0.202	0.399	0.498	0.197	0.391	0.492	0.012	0.007			
		(0.019)	(0.025)	(0.022)	(0.021)	(0.031)	(0.025)	(0.032)	(0.020)	95.3	98.8	94.0
		100	100	100	100	100	100	0.1	0.2			
	100	0.200	0.400	0.500	0.198	0.396	0.497	0.005	0.003			
		(0.008)	(0.011)	(0.010)	(0.009)	(0.014)	(0.011)	(0.013)	(0.009)	94.6	99.5	93.9
		100	100	100	100	100	100	0.3	0.4			
	200	0.200	0.400	0.500	0.199	0.397	0.497	0.004	0.002			
		(0.006)	(0.008)	(0.007)	(0.006)	(0.010)	(0.008)	(0.009)	(0.006)	94.0	99.7	93.3
		100	100	100	100	100	100	0.4	0.5			

Table C-3: Estimators obtained from $S = 1000$ simulations of model (4.2), for various values of N and T. Data are generated by using the Clayton copula with $\rho = 0.5$ and $p = 1$. Model (4.2) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	*AIC*	*BIC*	*QIC*
20	100	0.202	0.393	0.493	0.200	0.384	0.488	0.015	0.006			
		(0.019)	(0.043)	(0.034)	(0.020)	(0.051)	(0.037)	(0.044)	(0.023)	92.9	99.3	94.8
		100	100	100	100	99.9	100	0.1	0.2			
	200	0.201	0.397	0.496	0.199	0.391	0.492	0.010	0.005			
		(0.013)	(0.031)	(0.024)	(0.014)	(0.036)	(0.026)	(0.031)	(0.016)	93.6	99.7	95.8
		100	100	100	100	100	100	0.1	0.2			
100	10	0.233	0.364	0.460	0.223	0.349	0.450	0.027	0.011			
		(0.064)	(0.065)	(0.067)	(0.065)	(0.077)	(0.074)	(0.070)	(0.044)	84.9	87.2	87.4
		61.1	88.0	95.5	34.7	65.3	86.0	0.1	0.1			
	20	0.222	0.375	0.476	0.216	0.365	0.469	0.016	0.008			
		(0.045)	(0.048)	(0.049)	(0.046)	(0.057)	(0.054)	(0.050)	(0.030)	77.6	90.1	92.3
		99.3	99.7	100	98.8	99.2	100	0.7	0.2			
	100	0.203	0.393	0.493	0.201	0.389	0.489	0.008	0.004			
		(0.019)	(0.021)	(0.022)	(0.020)	(0.026)	(0.024)	(0.022)	(0.013)	81.2	90.7	93.2
		100	100	100	100	100	100	0.2	0.2			
	200	0.201	0.397	0.497	0.199	0.393	0.494	0.006	0.003			
		(0.014)	(0.015)	(0.015)	(0.014)	(0.018)	(0.017)	(0.015)	(0.009)	79.0	92.7	93.9
		100	100	100	100	100	100	0.4	0.5			

Table C-4: Estimators obtained from $S = 1000$ simulations of model (4.6), for various values of N and T. Data are generated by using the Gaussian copula with $\rho = 0.2$ and $p = 1$. Model (4.6) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	*AIC*	*BIC*	*QIC*
20	100	0.205	0.401	0.496	0.207	0.400	0.498	0.002	-0.004	81.2	97.2	85.2
		(0.047)	(0.019)	(0.027)	(0.051)	(0.033)	(0.033)	(0.035)	(0.028)			
		94.3	100	100	91.7	100	100	0.7	0.2			
	200	0.201	0.400	0.499	0.202	0.399	0.499	0.001	-0.001	81.2	98.7	85.5
		(0.033)	(0.013)	(0.019)	(0.036)	(0.023)	(0.023)	(0.025)	(0.020)			
		100	100	100	100	100	100	0.3	0.4			
100	10	0.239	0.396	0.479	0.240	0.386	0.477	0.016	-0.003	53.5	57.6	61.5
		(0.124)	(0.033)	(0.047)	(0.122)	(0.052)	(0.052)	(0.056)	(0.045)			
		17.1	100	99.9	12.7	97.5	99.3	0.9	0.2			
	20	0.221	0.399	0.490	0.223	0.393	0.489	0.008	-0.003	61.5	66.6	74.3
		(0.089)	(0.021)	(0.033)	(0.089)	(0.039)	(0.039)	(0.043)	(0.033)			
		39.4	100	100	37.9	100	100	1.4	0.9			
	100	0.209	0.399	0.497	0.209	0.399	0.498	0.000	-0.002	57.5	83.4	83.9
		(0.038)	(0.009)	(0.014)	(0.039)	(0.018)	(0.018)	(0.020)	(0.015)			
		99.5	100	100	99.4	100	100	0.6	0.1			
	200	0.204	0.400	0.498	0.204	0.400	0.499	0.000	0.000	59.3	87.8	85.9
		(0.027)	(0.006)	(0.010)	(0.028)	(0.013)	(0.013)	(0.014)	(0.010)			
		100	100	100	100	100	100	0.5	0.8			

Table C-5: Estimators obtained from $S = 1000$ simulations of model (4.6), for various values of N and T. Data are generated by using the Gaussian copula with $\rho = 0$ and $p = 1$. Model (4.6) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	*AIC*	*BIC*	*QIC*
20	100	0.202	0.401	0.498	0.203	0.400	0.498	0.002	-0.001	86.9	98.9	84.8
		(0.034)	(0.018)	(0.024)	(0.038)	(0.031)	(0.029)	(0.033)	(0.026)			
		99.9	100	100	99.9	100	100	0.4	0.5			
	200	0.202	0.400	0.498	0.202	0.400	0.499	0.000	-0.001	87.3	99.7	87.0
		(0.024)	(0.013)	(0.017)	(0.027)	(0.022)	(0.021)	(0.023)	(0.018)			
		100	100	100	100	100	100	0.3	0.5			
100	10	0.206	0.401	0.496	0.206	0.398	0.495	0.005	-0.001	88.2	90.3	77.8
		(0.049)	(0.026)	(0.029)	(0.050)	(0.038)	(0.037)	(0.041)	(0.030)			
		70.4	100	100	55.3	99.9	100	0.3	0.2			
	20	0.202	0.400	0.499	0.202	0.400	0.499	0.001	0.000	87.6	94.1	82.3
		(0.035)	(0.017)	(0.021)	(0.036)	(0.026)	(0.026)	(0.028)	(0.022)			
		99.3	100	100	98.8	100	100	0.7	0.5			
	100	0.201	0.400	0.500	0.201	0.400	0.500	0.000	-0.001	86.5	98.8	86.6
		(0.016)	(0.008)	(0.009)	(0.017)	(0.012)	(0.012)	(0.013)	(0.010)			
		100	100	100	100	100	100	0.5	0.4			
	200	0.200	0.400	0.500	0.200	0.400	0.500	0.000	0.000	85.2	99.3	85.1
		(0.012)	(0.005)	(0.007)	(0.012)	(0.008)	(0.008)	(0.009)	(0.007)			
		100	100	100	100	100	100	0.4	0.6			

Table C-6: Estimators obtained from $S = 1000$ simulations of model (4.6), for various values of N and T. Data are generated by using the Clayton copula with $\rho = 0.5$ and $p = 1$. Model (4.6) is also fitted using $p = 2$ to check the performance of various information criteria (IC). We use AIC, BIC and QIC.

Dim.		$p=1$			$p=2$					IC (%)		
N	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	AIC	BIC	QIC
20	100	0.208	0.403	0.492	0.211	0.401	0.494	0.004	-0.005	66.6	91.3	83.8
		(0.060)	(0.021)	(0.034)	(0.064)	(0.036)	(0.042)	(0.041)	(0.036)			
		80	100	100	74.9	100	100	0.9	0.3			
	200	0.203	0.401	0.497	0.204	0.399	0.498	0.002	-0.002	66.6	93.9	84.3
		(0.042)	(0.015)	(0.024)	(0.046)	(0.026)	(0.030)	(0.029)	(0.025)			
		97.5	100	100	96.1	100	100	0.4	0.5			
100	10	0.297	0.389	0.448	0.299	0.368	0.448	0.032	-0.010	34.8	37.5	57.2
		(0.166)	(0.039)	(0.067)	(0.162)	(0.071)	(0.074)	(0.081)	(0.062)			
		17.1	99.1	94.0	12.9	74.4	84.5	0.8	0.4			
	20	0.255	0.398	0.469	0.259	0.396	0.478	0.007	-0.016	31.5	40.5	71.6
		(0.126)	(0.025)	(0.049)	(0.128)	(0.057)	(0.060)	(0.065)	(0.048)			
		26.0	100	100	25.5	99.8	100	1.2	0.2			
	100	0.214	0.400	0.493	0.216	0.400	0.495	0.002	-0.004	28.7	52.1	83.1
		(0.057)	(0.011)	(0.022)	(0.059)	(0.028)	(0.029)	(0.032)	(0.023)			
		84.0	100	100	81.5	100	100	0.6	0.3			
	200	0.208	0.399	0.497	0.209	0.399	0.498	0.001	-0.002	32.0	55.7	81.2
		(0.040)	(0.008)	(0.015)	(0.042)	(0.020)	(0.021)	(0.023)	(0.017)			
		99.0	100	100	97.7	100	100	1.1	0.6			

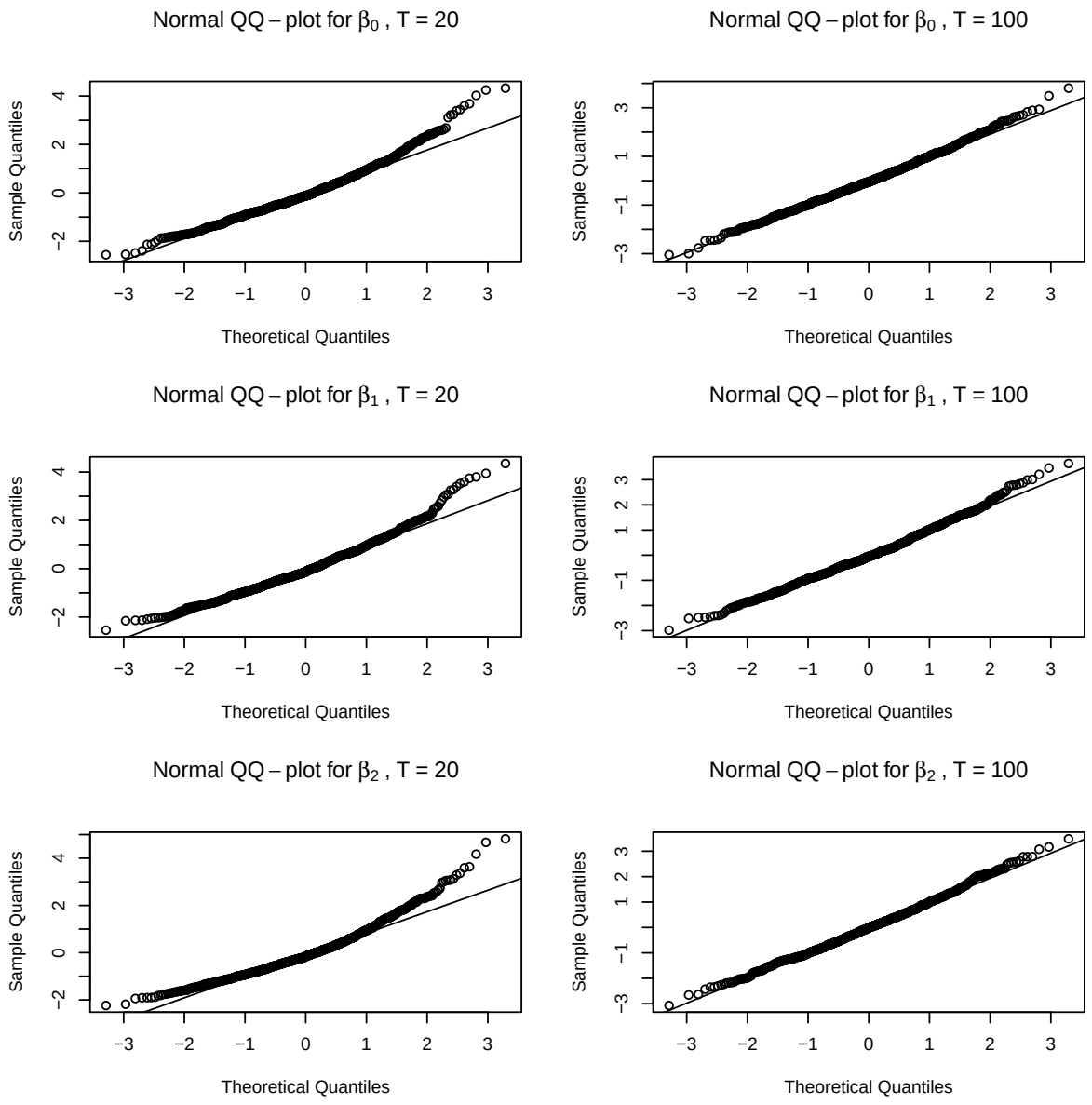

Figure C-5: QQ-plots for the log-linear model, Gaussian copula with $\rho = 0.5$, $N = 100$. Left: $T = 20$. Right: $T = 100$.

Bibliography

Ahmad, A. (2016). *Contributions à l'éconemétrie des séries temporelles à valeurs entières*. Ph. D. thesis, University Charles De Gaulle-Lille III, France.

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

Al-Osh, M. and A. A. Alzaid (1987). First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis 8*, 261–275.

Alzaid, A. and M. Al-Osh (1990). An integer-valued pth-order autoregressive structure (INAR (p)) process. *Journal of Applied Probability*, 314–324.

Andreassen, C. M. (2013). *Models and inference for correlated count data*. Ph. D. thesis, Aaarhus University, Denmark.

Andrews, D. W. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory 4*, 458–467.

Basawa, I. V. and B. L. S. Prakasa Rao (1980). *Statistical Inference for Stochastic Processes*. Academic Press, Inc., London-New York.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons.

Chen, X., Y. Chen, and P. Xiao (2013). The impact of sampling and network topology on the estimation of social intercorrelations. *Journal of Marketing Research 50*, 95–110.

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

Clark, N. J., M. S. Kaiser, and P. M. Dixon (2018). A spatially correlated auto-regressive model for count data. *arXiv preprint arXiv:1805.08323*.

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

Cui, Y. and Q. Zheng (2017). Conditional maximum likelihood estimation for a class of observation-driven time series models for count data. *Statistics & Probability Letters 123*, 193–201.

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

Davis, R. A., S. H. Holan, R. Lund, and N. Ravishanker (Eds.) (2016). *Handbook of Discrete-Valued Time Series*. London: Chapman & Hall/CRC.

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

Debaly, Z. M. and L. Truquet (2019). Stationarity and moment properties of some multivariate count autoregressions. *arXiv preprint arXiv:1909.11392*.

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

Douc, R., K. Fokianos, and E. Moulines (2017). Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electronic Journal of Statistics 11*, 2707–2740.

Doukhan, P. (1994). *Mixing*, Volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Doukhan, P., K. Fokianos, and D. Tjøstheim (2012). On weak dependence conditions for Poisson autoregressions. *Statistics & Probability Letters 82*, 942–948. with a correction in Vol. 83, pp. 1926-1927.

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

Fokianos, K. (2021). Multivariate count time series modelling. *arXiv preprint arXiv:2103.08028*.

Fokianos, K. and B. Kedem (2004). Partial likelihood inference for time series following generalized linear models. *Journal of Time Series Analysis 25*, 173–197.

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

Fokianos, K., B. Støve, D. Tjøstheim, and P. Doukhan (2020). Multivariate count autoregression. *Bernoulli 26*, 471–499.

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

Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *Astin Bull. 37*, 475–515.

Hall, P. and C. C. Heyde (1980). *Martingale Limit Theory and its Application*. Academic Press, Inc., New York-London.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson model. Technical Report MPRA Paper 8113, University Library of Munich, Germany. Available at `http://mpra.ub.uni-muenchen.de/8113/`.

Heinen, A. and E. Rengifo (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance 14*, 564 – 583.

Heyde, C. C. (1997). *Quasi-likelihood and its Application. A General Approach to Optimal Parameter Estimation*. Springer Series in Statistics. Springer-Verlag, New York.

Kedem, B. and K. Fokianos (2002). *Regression Models for Time Series Analysis*. John Wiley & Sons, Hoboken, NJ.

Knight, M., K. Leeming, G. Nason, and M. Nunes (2020). Generalized network autoregressive processes and the GNAR package. *Journal of Statistical Software 96*, 1–36.

Kolaczyk, E. D. and G. Csárdi (2014). *Statistical Analysis of Network Data with R*, Volume 65. Springer.

Latour, A. (1997). The multivariate GINAR (p) process. *Advances in Applied Probability 29*, 228–248.

Lee, Y., S. Lee, and D. Tjøstheim (2018). Asymptotic normality and parameter change test for bivariate Poisson INGARCH models. *TEST 27*, 52–69.

Liu, H. (2012). *Some models for time series of counts*. Ph. D. thesis, Columbia University, USA.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

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

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

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics 57*, 120–125.

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

Pedeli, X. and D. Karlis (2013a). On composite likelihood estimation of a multivariate INAR (1) model. *Journal of Time Series Analysis 34*, 206–220.

Pedeli, X. and D. Karlis (2013b). Some properties of multivariate INAR (1) processes. *Computational Statistics & Data Analysis 67*, 213–225.

Pötscher, B. M. and I. R. Prucha (1997). *Dynamic Nonlinear Econometric Models*. Springer-Verlag, Berlin. Asymptotic theory.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America 42*, 43–47.

Seber, G. A. F. (2008). *A Matrix Handbook for Statisticians*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Shiryaev, A. N. (2016). *Probability. 1* (Third ed.), Volume 95. Springer, New York.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Taniguchi, M. and Y. Kakizawa (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer Series in Statistics. Springer-Verlag, New York.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Wang, C., H. Liu, J.-F. Yao, R. A. Davis, and W. K. Li (2014). Self-excited threshold Poisson autoregression. *Journal of the American Statistical Association 109*, 777–787.

Wang, Y. J. and G. Y. Wong (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association 82*, 8–19.

Wasserman, S., K. Faust, et al. (1994). *Social Network Analysis: Methods and Applications*, Volume 8. Cambridge University Press.

Weiß, C. H. (2018). *An Introduction to Discrete-valued Time Series*. John Wiley & Sons.

Woodard, D. W., D. S. Matteson, and S. G. Henderson (2011). Stationarity of count-valued and nonlinear time series models. *Electronic Journal of Statistics 5*, 800–828.

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

Zeger, S. L. and K.-Y. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121–130.

Zhou, J., D. Li, R. Pan, and H. Wang (2020). Network GARCH model. *Statistica Sinica 30*, 1–18.

Zhu, X. and R. Pan (2020). Grouped network vector autoregression. *Statistica Sinica 30*, 1437–1462.

Zhu, X., R. Pan, G. Li, Y. Liu, and H. Wang (2017). Network vector autoregression. *The Annals of Statistics 45*, 1096–1123.

Zhu, X., W. Wang, H. Wang, and W. K. Härdle (2019). Network quantile autoregression. *Journal of Econometrics 212*, 345–358.

Chapter 5

Concluding remarks

In conclusion, we give some insight of future direction of study. We start from Chapter 3. First of all, we focus on the probabilistic properties. Although the uniqueness of the stationary distribution for the discrete valued processes are proved by using Markov chain theory, the rate of convergence to the limiting distribution is still unanswered. From the point of view of modelling improvements, an interesting extension could be achieved by considering a Markov chain of order greater than 1 which is able to define a model with several lags besides the first. As far as inferential model comparison is concerned, methods based on penalized likelihood, i.e., AIC and BIC, are adopted to compare the performance across various models, in terms of fitting and prediction. Nevertheless, theory and methods for model selection represent an important open issue, which need to be better investigated. Finally, in line with the recent theory developed for some multivariate discrete-valued processes, the specification of a unified framework for modelling multivariate discrete-valued time series may represent an interesting and challenging generalization.

For what concerns Chapter 4, is it worth to specify that throughout the paper the network has been assumed as nonrandom; this structure may be suitable for some fields (social network in short time periods, spatial borders) but it might be unrealistic for other applications (like epidemics). Therefore, a challenging and useful extension of the Network Autoregression (NAR) models either for continuous and discrete responses would be estimating a random adjacency matrix, and then casting it into the time series model. However, strong problems of model identifiability could raise as well as curse of dimensionality difficulties related to the contemporaneous estimations of several parameters. A second extension of crucial importance is related to the estimation. As the network dimension grows the QMLE obtained from the independence quasi-likelihood might be a poor estimator of the "true" parameters. A more suitable estimation procedure, for example using the generalized estimating equation theory, might be of interest for future researches.

List of Figures

List of Tables