JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **45**: 584–612 (2024)
Published online 19 December 2023 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12728

ORIGINAL ARTICLE

COUNT NETWORK AUTOREGRESSION

MIRKO ARMILLOTTA[a,b] AND KONSTANTINOS FOKIANOS[c]
[a]*Department of Econometrics and Data Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*
[b]*Tinbergen Institute, Amsterdam, The Netherlands*
[c]*Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus*

We consider network autoregressive models for count data with a non-random neighborhood structure. The main methodological contribution is the development of conditions that guarantee stability and valid statistical inference for such models. 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 article is complemented by simulation results and a data example.

Received 14 May 2022; Accepted 15 November 2023

Keywords: Generalized linear models; increasing dimension; link function; multi-variate count time series; quasi-likelihood.

MOS subject classification: 62M10.

1. INTRODUCTION

The vast availability of integer-valued data, emerging from several real-world applications, has motivated the growth of a large body of literature for modeling and inference of count time series processes. For comprehensive surveys, see Kedem and Fokianos (2002), Weiß (2018), Davis *et al.* (2021), among others. The aim of this contribution is to develop a statistical framework for network count time series which are simply multi-variate time series equipped with a neighborhood structure. Consider the vector which consists of all node measurements at some time t. This is going to be the response vector we will be studying and we will assume that its evolution is influenced not only by past observations but also by its neighbors. We consider such processes assuming that their neighborhood structure is known. We deal with a multi-variate problem whose main challenge is that the response vector is high-dimensional and therefore we study, in detail, this case as we explain below.

1.1. Related Work

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 particularly useful as we will be developing theory for count 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 also be 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. Related

*Correspondence to: Konstantinos Fokianos, Department of Mathematics and Statistics, University of Cyprus, University House "Anastasios G. Leventis" 1 Panepistimiou Avenue 2109 Aglantzia, P.O. Box 20537, 1678 Nicosia, Cyprus.
Email: fokianos@ucy.ac.cy

work includes Fokianos *et al.* (2009) and Fokianos and Tjøstheim (2011) who develop properties and estimation for a class of linear and log-linear count time series models. Further related contributions have 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. In addition, Douc *et al.* (2013, 2017), Dunsmuir (2016), Davis and Liu (2016), Cui and Zheng (2017), and more recently Armillotta *et al.* (2022), among others, provide further generalizations of observation-driven models leaning on general distribution functions or one-parameter exponential family of distributions. 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).

Studies of multi-variate INAR models include those of Latour (1997), Pedeli and Karlis (2011, 2013a, b), among others. Theory and inference for multi-variate 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 multi-variate extension of the linear and log-linear Poisson autoregression model, by employing a copula-based construction for the joint distribution of the counts. The 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, pp. 507-508). They propose a plausible data generating process (DGP) which preserves, marginally, Poisson processes' properties, conditional on the past. Further details are given by the recent review of Fokianos (2021).

1.2. Network Time Series

Multi-variate observation-driven count time series models are useful for modeling 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 multi-variate time series process has attracted considerable attention over the last years. In an unpublished work, Knight *et al.* (2016) defined multi-variate continuous time series coupled with a network structure as network time series. Furthermore these authors proposed methodology for the analysis of such data. Such approach has been originally proposed in the context of spatiotemporal data analysis, referred to as Space-Time Autoregressive Moving Average (STARMA) models; Cliff and Ord (1975), Martin and Oeppen (1975) and Pfeifer and Deutrch (1980), among many others. In general, any stream of data for a sample of units whose relations can be modeled through an adjacency matrix (neighborhood structure), adhere to statistical techniques developed in this article. Zhu *et al.* (2017) have discussed a similar model, called Network Autoregressive model (NAR), which is an autoregressive model for continuous valued 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. More precisely, it is assumed that $N \to \infty$ and $T_N \to \infty$, i.e., the temporal sample size is assumed to depend on N. The regime (a) corresponds to standard asymptotic inference in time series analysis. However, in network analysis it is important to understand the behavior of the process when the network's dimension grows. This is a relevant problem in fields where typically the network is large, see, for example, social networks in Wasserman *et al.* (1994). It is also essential to have stability conditions for large network structures, so that proper time series inference can be advanced; those problems motivate study of asymptotics under regime (b). 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) but for the case of $T \to \infty$ and fixed network dimension N. Under the standard asymptotic regime (a), 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 neighbors of the network. Moreover, the same authors provide R software (package GNAR) for fitting such models.

DOI: 10.1111/jtsa.12728 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

1.3. Our Contribution

Integer-valued 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 (number of posts, number of likes, counts of digit employed in comments, etc). Another typical field of application is related to the number of cases in epidemic models for studying the spread of infection diseases in a population; this is even more important in the current COVID-19 pandemic outbreak. Recently, an application of this type which employs a model similar to the NAR with count data has been suggested by Bracher and Held (2022). Therefore, the extension of the NAR model to multi-variate 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 studying in detail 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 assumption that the innovations are i.i.d. 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 an alternative approach to modeling multi-variate counts in continuous time see Veraart (2019), Eyjolfsson and Tjøstheim (2023), and Fang *et al.* (2021) for a network model employing Hawkes processes which are related to the linear and log-linear model we will be studying. Indeed those models are obtained after suitable discretization of the corresponding continuous time process. However our proposal imposes a specific DGP, does not assume homogeneity across the network and the condition required for obtaining good large sample properties of the QMLE are quite different than those assumed by Fang *et al.* (2021).

For the continuous-valued case, Zhu *et al.* (2017) employed ordinary least square (OLS) estimation combined with specific properties imposed on the adjacency matrix for the estimation of unknown model parameters. However, this method is not applicable to general time series models. In the case we study, estimation is carried out by using quasi-likelihood methods; see Heyde (1997), for example. When the network dimension N is fixed and $T \to \infty$, standard results for Quasi Maximum Likelihood Estimation (QMLE) from multi-variate count autoregressions, as developed by Fokianos *et al.* (2020), carry over to the case of PNAR models. When the network dimension is increasing, the asymptotic properties of the estimators would rely on the ergodicity of a stationary random process $\{\mathbf{Y}_t : t \in \mathbb{Z}\}$ with $N \to \infty$. However, there exists no widely accepted definition for stationarity of a process with infinite dimension. Consequently no ergodicity results are available for processes with $N \to \infty$ and standard time series results concerning convergence of sample means do not carry over to the increasing dimension case. In the present contribution, this problem is bypassed by providing an alternative proof, based on the laws of large numbers for L^p-NED processes of Andrews (1988). Our method employs the working definition of stationarity of Zhu *et al.* (2017, Def. 1) for processes of increasing dimension. All these developments are crucial to a thorough study of QMLE under the double regime asymptotics we consider. Finally, we are addressing several other related problem, including estimation of contemporaneous dependence and improving the efficiency of the QMLE.

1.4. Outline

The article is organized as follows: Section 2 discusses the PNAR(p) model specification for the linear and the log-linear case, with lag order p, and the related stability properties. In Section 3, quasi-likelihood inference is established, showing consistency and asymptotic normality of the QMLE for the two types of asymptotics (a) and (b). Section 4 discusses the results of a simulation study and an application on real data. The article concludes

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

with an Appendix containing the proofs of Theorem 1 and Lemma 1 and 2. All the other proofs are included in the Supplement SM together with additional results.

1.4.1. Notation

We denote $|\mathbf{x}|_r = \left(\sum_{j=1}^d \left|x_j\right|^r\right)^{1/r}$ the l^r-norm of a d-dimensional vector $\mathbf{x}$. If $r = \infty$, $|\mathbf{x}|_\infty = \max_{1\le j\le d}|x_j|$. Let $\|\mathbf{X}\|_r = \left(\sum_{j=1}^d \mathrm{E}(|X_j|^r)\right)^{1/r}$ 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. Define $\lambda_{\max}(\mathbf{M})$ the largest absolute eigenvalue of a symmetric matrix $\mathbf{M}$. Define $|\mathbf{x}|_v = \left(|x_1|, \dots, |x_d|\right)'$, $|\mathbf{A}|_v = \left(\left|a_{i,j}\right|\right)_{(i,j)}$ and $\|\mathbf{X}\|_v = \left(\mathrm{E}\left|X_1\right|, \dots, \mathrm{E}\left|X_d\right|\right)'$ the elementwise l^1-norm for vectors, matrices and random vectors respectively. Moreover, denote by $\prec$ a partial order relation on $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\mathbf{x} \prec \mathbf{y}$ means $x_i \le y_i$ for $i = 1, \dots, d$. For a d-dimensional vector $\mathbf{x}$, with $d \to \infty$, set the following compact notation $\sup_{1\le i<\infty} x_i = \sup_{i\ge 1} x_i$. The notations C_r and D_r denote a constant which depend on r, where $r \in \mathbb{N}$. In particular C denotes a generic constant. Finally, throughout the article the notation $\left\{N, T_N\right\} \to \infty$ will be used as a shorthand for $N \to \infty$ and $T_N \to \infty$, where the temporal size T is assumed to depend on the network dimension N.

2. STABILITY RESULTS FOR COUNT NETWORK TIME SERIES

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 $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N\times N}$, i.e. $a_{ij} = 1$ provided that there exists 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 non-random, by this we mean that the network is known with fixed edges; see also Zhu *et al.* (2017). Self-relationships are not allowed, i.e. $a_{ii} = 0$ for any $i = 1, \dots, N$; this is a typical assumption, and it is reasonable for various real situations, e.g. social networks, where users do not follow themselves; see Wasserman *et al.* (1994) and Kolaczyk and Csárdi (2014). Define a 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 $\left\{Y_{i,t}\right\}$ for $i = 1, \dots, N$ over time $t = 1, \dots, T$.

Here, 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 $\left\{\mathbf{Y}_t = (Y_{i,t},\ i = 1, 2, \dots, N,\ t = 0, 1, 2 \dots, T)\right\}$ an N-dimensional vector of count time series with $\left\{\lambda_t = (\lambda_{i,t},\ i = 1, 2, \dots, N,\ t = 1, 2, \dots, T)\right\}$ 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 $\lambda_t = \mathrm{E}(\mathbf{Y}_t|\mathcal{F}_{t-1})$.

2.1. Linear PNAR(1) Model

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

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

where $\beta_0, \beta_1, \beta_2 \ge 0$ and $n_i = \sum_{j\ne i} a_{ij}$ is the out-degree, i.e the total number of nodes which i has an edge with. From the left-hand side equation of (1), we observe that the process $Y_{i,t}$ is assumed to be marginally Poisson, conditionally to the past. We call (1) linear Poisson network autoregression of order 1, abbreviated by PNAR(1).

Model (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 $j \ne i$ 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 $n_i^{-1}\sum_{j=1}^{N} a_{ij}Y_{j,t-1}$. The coefficient β_2 is called momentum effect because it provides a weight for the impact of past count $Y_{i,t-1}$. This interpretation is in line with the Gaussian NAR as discussed by Zhu *et al.* (2017) for the case of continuous variables.

Equation (1) does not include information about the joint dependence structure of the PNAR(1) model.

Then the goal is to introduce a multi-variate random vector, at each time point t, whose each component follow marginally the Poisson distribution (conditionally to the past) but there exists among them arbitrary correlation. In a recent work, Fokianos *et al.* (2020) defined such a distribution in terms of a DGP specified by an algorithm which generates a random vector whose dependence among their components is introduced by imposing a copula on the waiting times of a Poisson process; see also (Fokianos, 2021, p. 4). In this way, we can, define the multi-variate copula Poisson distribution with parameter, say $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N)^T$, and denote it by $MCP(\boldsymbol{\lambda})$, as an N-dimensional random vector whose components are marginally Poisson distributed with mean λ_i, $i = 1, 2, \dots, N$ and whose structure of dependence is modeled through the copula $C(\dots)$ on their associated exponential waiting times random variables. It is then convenient to rewrite (1) in vectorial form, following Fokianos *et al.* (2020),

$$\mathbf{Y}_t|\mathcal{F}_{t-1} \sim \mathrm{MCP}(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \boldsymbol{\beta}_0 + \mathbf{G}\mathbf{Y}_{t-1}, \tag{2}$$

where $\boldsymbol{\beta}_0 = \beta_0\mathbf{1}_N \in \mathbb{R}^N$, with $\mathbf{1} = (1, 1, \dots, 1)^T \in \mathbb{R}^N$, and the matrix $\mathbf{G} = \beta_1\mathbf{W} + \beta_2\mathbf{I}_N$, where $\mathbf{W} = \mathrm{diag}\left\{n_1^{-1}, \dots, n_N^{-1}\right\}\mathbf{A}$ is the row-normalized adjacency matrix, with $\mathbf{A} = (a_{ij})$, so $\mathbf{w}_i = (a_{ij}/n_i, j = 1, \dots, N)^T \in \mathbb{R}^N$ is the ith row vector of the matrix $\mathbf{W}$, satisfying $|||\mathbf{W}|||_\infty = 1$, and $\mathbf{I}_N$ is the $N \times N$ identity matrix. In general, the weights $\mathbf{w}_i$ can be chosen arbitrarily as long as $|||\mathbf{W}|||_\infty = 1$ is satisfied. To obtain insight for the DGP, as introduced by (2), consider a set of values $(\beta_0, \beta_1, \beta_2)^T$ and a starting vector $\boldsymbol{\lambda}_0 = (\lambda_{1,0}, \dots, \lambda_{N,0})^T$,

1. Let $\mathbf{U}_l = (U_{1,l}, \dots, U_{N,l})$, for $l = 1, \dots, K$ a sample from a N-dimensional copula $C(u_1, \dots, u_N)$, where $U_{i,l}$ follows a Uniform(0,1) distribution, for $i = 1, \dots, N$.
2. The transformation $X_{i,l} = -\log U_{i,l}/\lambda_{i,0}$ follows the exponential distribution with parameter $\lambda_{i,0}$, for $i = 1, \dots, N$.
3. If $X_{i,1} > 1$, then $Y_{i,0} = 0$, otherwise $Y_{i,0} = \max\left\{k \in [1, K] : \sum_{l=1}^{k} X_{i,l} \le 1\right\}$, by taking K large enough. Then, $Y_{i,0}|\boldsymbol{\lambda}_0 \sim \mathrm{Poisson}(\lambda_{i,0})$, for $i = 1, \dots, N$. So, $\mathbf{Y}_0 = (Y_{1,0}, \dots, Y_{N,0})$ is a set of (conditionally) marginal Poisson processes with mean $\boldsymbol{\lambda}_0$.
4. By using the model (2), $\boldsymbol{\lambda}_1$ is obtained.
5. Return back to step 1 to obtain $\mathbf{Y}_1$, and so on.

In practical applications the sample size K should be a large value, e.g. $K = 1000$; its value clearly depends, in general, on the magnitude of observed data. Moreover, the copula construction $C(\dots)$ will depend on one or more unknown parameters, say ρ, which capture the contemporaneous correlation among the variables.

The previous algorithm generates a sample of multi-variate counts for practical simulations. In principle, the algorithm simulates realizations of a stochastic process $\left\{\mathbf{Y}_t; t \in \mathbb{Z}\right\}$, i.e. for all integers. Accordingly, $\boldsymbol{\lambda}_0$ is not a fixed vector but $\boldsymbol{\lambda}_0 = \boldsymbol{\beta}_0 + \mathbf{G}\mathbf{Y}_{-1}$, being a function of $\mathcal{F}_{-1}$, then $Y_{i,0}|\boldsymbol{\lambda}_0 \sim \mathrm{Poisson}(\lambda_{i,0})$ is equivalent to say $Y_{i,0}|\mathcal{F}_{-1} \sim \mathrm{Poisson}(\lambda_{i,0})$. The same happens for $\boldsymbol{\lambda}_{-1}$ ad so on. Then, the DGP generates $Y_{i,t}$ being conditionally marginally Poisson for all $t \in \mathbb{Z}$.

The development of a multi-variate count time series model would be based on specification of a joint distribution, so that the standard likelihood inference and testing procedures can be developed. Although several alternatives have been proposed in the literature, see the review in Fokianos (2021, sec. 2), the choice of a suitable multi-variate version of the conditional Poisson probability mass function (p.m.f) is a challenging problem. In fact, multi-variate Poisson-type p.m.f have usually complicated closed form and the associated likelihood inference is theoretically and computationally cumbersome. Furthermore, in many cases, the

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

available multi-variate Poisson-type p.m.f. implicitly imply restrictions on models with limited use in applications (e.g. covariances always positive, constant pairwise correlations). In this article the joint distribution of the vector $\{\mathbf{Y}_t\}$ is constructed by following the copula approach described above. The proposed DGP ensures that all marginal distributions of $Y_{i,t}$ are univariate Poisson, conditionally to the past, as described in (1), while it introduces an arbitrary dependence among them in a flexible and general way by the copula construction. See Inouye *et al.* (2017) and Fokianos (2021) for a discussion on the choice of multi-variate count distributions and several alternatives. Further results regarding the empirical properties of model (2) are discussed in Section S-1.2 of Supplement SM.

We choose the conditional multi-variate copula Poisson distribution for its simplicity and because it is a natural distributional assumption for counting number of events over a time period. However, any multi-variate count distribution whose mean is modeled through (1) and possesses moments up to an appropriate order fits the QMLE methodology which employs (10) to derive consistent and asymptotically normally distributed estimators. In fact, theory and applications can be extended to other count distributions. By exploiting the same copula construction and modifying suitably the generation of exponential waiting times, we can define a conditional copula multi-variate Negative Binomial distribution, and more generally a conditional copula mixed Poisson distribution; see Fokianos *et al.* (2020, p. 474). A complete treatment of such extensions remains unexplored.

2.2. Linear PNAR(p) Model

More generally, we introduce and study an extension of model (1) by allowing $Y_{i,t}$ to depend on the last p lagged values. We call this the linear Poisson NAR(p) model and its defined analogously to (1) but with

$$\lambda_{i,t} = \beta_0 + \sum_{h=1}^{p} \beta_{1h} \left(n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j,t-h} \right) + \sum_{h=1}^{p} \beta_{2h} Y_{i,t-h}, \tag{3}$$

where $\beta_0, \beta_{1h}, \beta_{2h} \geq 0$ for all $h = 1 \ldots, p$. If $p = 1$, set $\beta_{11} = \beta_1$, $\beta_{22} = \beta_2$ to obtain (1). The joint conditional distribution of the vector $\mathbf{Y}_t$ is defined by means of the copula construction discussed in Section 2.1. Without loss of generality, we can set coefficients equal to zero if the parameter order is different in both terms of (3). Then (3) is rewritten as

$$\mathbf{Y}_t | \mathcal{F}_{t-1} \sim MCP(\lambda_t) \quad \lambda_t = \boldsymbol{\beta}_0 + \sum_{h=1}^{p} \mathbf{G}_h \mathbf{Y}_{t-h}, \tag{4}$$

where $\mathbf{G}_h = \beta_{1h}\mathbf{W} + \beta_{2h}\mathbf{I}_N$ for $h = 1, \ldots, p$ by recalling that $\mathbf{W} = \text{diag}\left\{n_1^{-1}, \ldots, n_N^{-1}\right\}\mathbf{A}$. The following result establishes sharp verifiable conditions for proving ergodicity, when N is fixed.

Proposition 1. *Consider model (4), with fixed N. Suppose that $\rho\left(\sum_{h=1}^{p} \mathbf{G}_h\right) < 1$. 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 $r \geq 1$.*

The result follows from Debaly and Truquet (2021, thm. 2). Similar results have been recently proved by Fokianos *et al.* (2020) when the lagged conditional mean λ_{t-1} is added as a feedback term in the model. Following these authors, we obtain the same results of Proposition 1 but under stronger conditions. For example, when $p = 1$, we will need to assume either $|||\mathbf{G}|||_1 < 1$ or $|||\mathbf{G}|||_2 < 1$ to obtain identical conclusions. Results about the first and second-order properties of model (3) are given in Section S-1 in Supplement SM; see also Fokianos *et al.* (2020, prop. 3.2).

Proposition 1 establishes the existence of the moments of the count process with fixed N, but this property is not guaranteed to hold when $N \to \infty$. The following results show that, is $N \to \infty$, the conclusions of Proposition 1 are still true.

Proposition 2. *Consider model (4) and $\sum_{h=1}^{p}(\beta_{1h} + \beta_{2h}) < 1$. Then, $\sup_{i \geq 1} \mathrm{E}|Y_{i,t}|^r \leq C_r < \infty$, for any $r \in \mathbb{N}$.*

DOI: 10.1111/jtsa.12728

In order to investigate the stability results of the process $\{\mathbf{Y}_t \in \mathbb{N}^N\}$ when the network size is diverging ($N \to \infty$) we employ the working definition of stationarity for increasing dimensional processes as discussed by Zhu *et al.* (2017, def. 1). The following result holds.

Theorem 1. *Consider model (4). Assume* $\sum_{h=1}^{p}(\beta_{1h} + \beta_{2h}) < 1$ *and* $N \to \infty$. *Then, there exists a unique strictly stationary solution* $\{\mathbf{Y}_t \in \mathbb{N}^N,\ t \in \mathbb{Z}\}$ *to the linear PNAR(p) model, with* $\sup_{i\geq 1} \mathrm{E}\left|Y_{i,t}\right|^r \leq C_r < \infty$, *for all* $r \geq 1$.

Theorem. 1 extends (Zhu *et al.*, 2017, thm.1). Although stronger than the conditions of Proposition 1, $\sum_{h=1}^{p}(\beta_{1h} + \beta_{2h}) < 1$ allows to prove stationarity for increasing network size N and the existence of moments of the process; moreover, it is more natural assumption than the condition $\rho(\sum_{h=1}^{p}\mathbf{G}_h) < 1$, and it complements the existing work for continuous valued models; Zhu *et al.* (2017). It is worth pointing out that the copula construction is not used in the proof of Theorem 1 (see also Theorem 2 for log-linear model). However, it is used in Section 4.1 where we report a simulation study. It is interesting though, that even under this setup, stability conditions are independent of the correlation structure of innovations; this is similar to the case of multi-variate ARMA models.

Remark 1. Models (4) implies that $Y_{i,t}$ are marginally Poisson distributed conditionally on the past of the process, $\mathcal{F}_{t-1}$. There is no any assumption about the marginal and joint unconditional distributions of the process. In general, the unconditional distribution of $\mathbf{Y}_t$ is unknown. However, from the results of Theorem 1 we can conclude that $\mathbf{Y}_t$ is a stationary Markov chain of order p so its (unconditional) distribution exists, is unique, does not depend on t and all its moments are uniformly bounded. Moreover, we derive explicitly the first two moments of such distribution (Section S-1 in Supplement SM).

Remark 2. A count GNAR(p) extension similar to the model introduced by Knight *et al.* (2020, eq. 1), for the standard asymptotic regime ($T \to \infty$), in the context of continuous-valued random variables, is included in the framework we consider. Such model adds an average neighbor impact for several stages of connections between the nodes of a given network. That is, $\mathcal{N}^{(r)}(i) = \mathcal{N}\left\{\mathcal{N}^{(r-1)}(i)\right\} / \left[\left\{\cup_{q=1}^{r-1}\mathcal{N}^{(q)}(i)\right\} \cup \{i\}\right]$, for $r = 2, 3, \ldots$ and $\mathcal{N}^{(1)}(i) = \mathcal{N}(\{i\})$, with $\mathcal{N}(\{i\}) = \{j \in \{1, \ldots, N\} : i \to j\}$ the set of neighbors of the node i. (So, e.g., $\mathcal{N}^{(2)}(i)$ describes the neighbors of the neighbors 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 $w_{i,j} = 1/\mathrm{card}(\mathcal{N}^{(r)}(i))$, $\mathrm{card}(\cdot)$ denotes the cardinality of a set and $I(\cdot)$ is the indicator function. Several M types of edges are allowed in the network. The Poisson GNAR(p) has the following formulation.

$$\lambda_{i,t} = \beta_0 + \sum_{h=1}^{p}\left(\sum_{m=1}^{M}\sum_{r=1}^{s_h}\beta_{1,h,r,m}\sum_{j\in\mathcal{N}_t^{(r)}(i)} w_{i,j,m}Y_{j,t-h} + \beta_{2,h}Y_{i,t-h}\right), \tag{5}$$

where s_h is the maximum stage of neighbor dependence for the time lag h and all the parameters of the model need to be non-negative. Model (5) can be included in the formulation (4) by setting $\mathbf{G}_h = \sum_{m=1}^{M}\sum_{r=1}^{s_h}\beta_{1,h,r,m}\mathbf{W}^{(r,m)} + \beta_{2,h}\mathbf{I}_N$. Since it holds that $\sum_{j\in\mathcal{N}^{(r)}(i)}\sum_{m=1}^{M} w_{i,j,m} = 1$, we have $\left|\left|\left|\sum_{m=1}^{M}\mathbf{W}^{(r,m)}\right|\right|\right|_\infty = 1$. Hence, the result of the present contribution, i.e., existence of the moments of the model, the related stability properties and the associated inferential results, under the standard asymptotic regime, apply to (5).

2.3. Log-linear PNAR models

Recall model (1). The network effect β_1 of model (1) is typically expected to be positive, see Chen *et al.* (2013), and the impact of $Y_{i,t-1}$ is positive, as well. Hence, positive constraints on the parameters are theoretically justifiable as well as practically sound. However, in order to allow a natural link to the GLM theory, McCullagh and Nelder (1989), and allowing the possibility to include covariates as well as real valued coefficients, we additionally

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

study the following log-linear model, see Fokianos and Tjøstheim (2011):

$$Y_{i,t}|\mathcal{F}_{t-1} \sim \text{Poisson}(\exp(\nu_{i,t})), \quad \nu_{i,t} = \beta_0 + \beta_1 n_i^{-1} \sum_{j=1}^{N} a_{ij} \log(1 + Y_{j,t-1}) + \beta_2 \log(1 + Y_{i,t-1}), \tag{6}$$

where $\nu_{i,t} = \log(\lambda_{i,t})$ for every $i = 1, \dots, N$. No parameters constraints are required for model (6) since $\nu_{i,t} \in \mathbb{R}$. Interpretation of all parameters is the same, as in the case of (1), but in the logarithmic scale. Again, the model can be rewritten in vectorial form, as in the case of model (2)

$$\mathbf{Y}_t|\mathcal{F}_{t-1} \sim \text{MCP}(\exp(\boldsymbol{\nu}_t)), \quad \boldsymbol{\nu}_t = \boldsymbol{\beta}_0 + \mathbf{G} \log(\mathbf{1}_N + \mathbf{Y}_{t-1}), \tag{7}$$

where $\text{MCP}(\exp(\boldsymbol{\nu}_t))$ is an N-dimensional copula conditional Poisson distribution, as above. Furthermore, it can be useful rewriting the model as follow.

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

where $\boldsymbol{\psi}_t = \log(\mathbf{1}_N + \mathbf{Y}_t) - \boldsymbol{\nu}_t$. By lemma A.1 in Fokianos and Tjøstheim (2011) $\text{E}(\boldsymbol{\psi}_t|\mathcal{F}_{t-1}) \to 0$ as $\boldsymbol{\nu}_t \to \infty$, so $\boldsymbol{\psi}_t$ is approximately martingale difference sequence (MDS). This means that the formulation of first two moments established for the linear model in Section S-1 in Supplement SM hold, approximately, for $\log(\mathbf{1}_N + \mathbf{Y}_t)$. We discuss empirical properties of the count process $\mathbf{Y}_t$ of model (6) in Section S-2.3 of the Supplement SM. Moreover, $\boldsymbol{\xi}_t = \mathbf{Y}_t - \exp(\boldsymbol{\nu}_t)$ is a MDS. We define the log-linear PNAR(p) by

$$\nu_{i,t} = \beta_0 + \sum_{h=1}^{p} \beta_{1h} \left(n_i^{-1} \sum_{j=1}^{N} a_{ij} \log(1 + Y_{j,t-h}) \right) + \sum_{h=1}^{p} \beta_{2h} \log(1 + Y_{i,t-h}), \tag{8}$$

using the same notation as before. Then

$$\mathbf{Y}_t|\mathcal{F}_{t-1} \sim \text{MCP}((\exp(\boldsymbol{\nu}_t)), \quad \boldsymbol{\nu}_t = \boldsymbol{\beta}_0 + \sum_{h=0}^{p} \mathbf{G}_h \log(\mathbf{1}_N + \mathbf{Y}_{t-h}), \tag{9}$$

where $\mathbf{G}_h = \beta_{1h}\mathbf{W} + \beta_{2h}\mathbf{I}_N$ for $h = 1, \dots, p$. The following results are complementing Propositions 1,2 and Theorem 1 proved for the case of log-linear model.

Proposition 3. *Consider model (9), with fixed N. Suppose that $\rho\left(\sum_{h=1}^{p} |\mathbf{G}_h|_v\right) < 1$. Then the process $\{\mathbf{Y}_t,\ t \in \mathbb{Z}\}$ is stationary and ergodic with $\text{E}|\mathbf{Y}_t|_1 < \infty$. Moreover, if $\left|\left|\left|\sum_{h=1}^{p} |\mathbf{G}_h|_v\right|\right|\right|_\infty < 1$, there exists some $\delta > 0$ such that $\text{E}[\exp(\delta |\mathbf{Y}_t|_1)] < \infty$ and $\text{E}[\exp(\delta |\boldsymbol{\nu}_t|_1)] < \infty$.*

The result follows from Debaly and Truquet (2019, thm. 5). Analogously to the linear model, we need to show the uniform boundedness of moments of the process and the stationarity of the model with increasing dimension. Since the noise $\boldsymbol{\psi}_t$ is approximately MDS, the following result is proved by employing approximate arguments.

Proposition 4. *Consider model (9) and $\left|\sum_{h=1}^{p}(\beta_{1h} + \beta_{2h})\right| < 1$. Then, $\sup_{i\geq 1} \text{E}|Y_{i,t}|^r \leq C_r < \infty$, and $\sup_{i\geq 1} \text{E}[\exp(r|\nu_{i,t}|)] \leq D_r < \infty$, for any $r \in \mathbb{N}$.*

Analogously to Theorem 1, a strict stationarity result for network of increasing order is given for the log-linear PNAR model (9).

Theorem 2. *Consider model (9). Assume $\sum_{h=1}^{p}(|\beta_{1h}| + |\beta_{2h}|) < 1$ and $N \to \infty$. Then, there exists a unique strictly stationary solution $\{\mathbf{Y}_t \in \mathbb{N}^N,\ t \in \mathbb{Z}\}$ to the log-linear PNAR model, with $\sup_{i\geq 1} \text{E}|Y_{i,t}|^r \leq C_r < \infty$, and $\sup_{i\geq 1} \text{E}[\exp(r|\nu_{i,t}|)] \leq D_r < \infty$, for all $r \geq 1$.*

DOI: 10.1111/jtsa.12728 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

Remark 3. For simplicity, model (4) has been defined without including covariates. But time-invariant positive covariates $\mathbf{Z} \in \mathbb{R}_+^d$ can be included without affecting the results of the present contribution, under suitable moments existence assumptions. This is a useful fact because we can consider available node-specific characteristics, for example. Moreover, the log-linear version (9) ensures the inclusion of covariates whose values belong to $\mathbb{R}^d$.

Remark 4. Analogous arguments made in Remark 2 for the linear model case hold true for the log-linear model (8) and a log-linear GNAR(p) can be advanced.

3. QUASI-LIKELIHOOD INFERENCE FOR INCREASING NETWORK SIZE

We develop inference for the unknown vector of parameters of models (4) and (9), denoted by $\boldsymbol{\theta} = (\beta_0, \beta_{11}, \dots, \beta_{1p}, \beta_{21}, \dots, \beta_{2p})^T \in \Lambda \subset \mathbb{R}^m$, where $m = 2p + 1$ and Λ is the parameter space. Full parametric likelihood inference requires specification of the conditional joint p.m.f., which is hard to obtain, because the exponential waiting times employed for steps 2 and 3 of the DGP algorithm are latent random variables. This implies that the imposed copula function cannot be used to obtain the full model likelihood. Nevertheless, the marginal conditional distributions of the DGP are well-defined quantities and can be employed for estimation of unknown parameters. Then, the estimation problem is approached by using the quasi maximum likelihood theory; see Wedderburn (1974) and Gourieroux *et al.* (1984) among others. Developing proofs of consistency and asymptotic normality of the QMLE, when $N \to \infty$ and $T_N \to \infty$, is the main goal of the present section. Define the conditional quasi log-likelihood function for the vector of unknown parameters $\boldsymbol{\theta}$ by

$$l_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{i=1}^{N} \left(Y_{i,t} \log \lambda_{i,t}(\boldsymbol{\theta}) - \lambda_{i,t}(\boldsymbol{\theta}) \right) \equiv \sum_{t=1}^{T}\sum_{i=1}^{N} l_{i,t}(\boldsymbol{\theta}), \tag{10}$$

which is the log-likelihood one would obtain if time series modeled in (4), or (9), are contemporaneously independent. Clearly such an approach does not require any specification/estimation of the copula structure $C(\dots, \rho)$ and its set of parameters ρ. Note that although the copula is not included in the maximization of the 'working' log-likelihood (10), the QMLE is not computed under the assumption of independence; this is easily seen by the form of the information matrix (15) below, which depends on the true conditional covariance matrix of the process $\mathbf{Y}_t$.

The quasi log-likelihood (10) allows computational simplifications and guarantees valid asymptotic properties of the estimator at the cost of a lower efficiency when compared top the full maximum likelihood estimator. In particular, (10) is a member of the one-parameter exponential family; then, even though we do not employ the true likelihood, Gourieroux *et al.* (1984, thm. 1-3) gives an indication that the resulting estimator will be consistent and asymptotically normal. Note that we study a different framework since both T, N are assumed to tend to infinity. Since $\mathbf{W}$ is a non-random sequence of matrices indexed by N, the specification of the asymptotic properties of the estimator deals with two diverging indexes, $N \to \infty$ and $T \to \infty$, allowing to establish a double-dimensional-type of converge, when both the temporal size and the network dimension grow together. Assuming that there exists a true vector of parameter, say $\boldsymbol{\theta}_0$, such that the mean model specification (4) (or equivalently (9)) is correct, regardless the true DGP, then we obtain a consistent and asymptotically normal estimator by maximizing the quasi log-likelihood (10). This is a novel result as most contributions in the literature deal either with the case $N = 1$ or N fixed; see previous references.

Consider the linear PNAR model (4). Denote by $\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}\in\Lambda} l_{NT}(\boldsymbol{\theta})$, the QMLE for $\boldsymbol{\theta}$. The score function for the linear model is given by

$$\begin{aligned} \mathbf{S}_{NT}(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{i=1}^{N} \left(\frac{Y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \left(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta}) \right) = \sum_{t=1}^{T} \mathbf{s}_{Nt}(\boldsymbol{\theta}) \,, \end{aligned} \tag{11}$$

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

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 (multiplied by -1) is given by

$$\mathbf{H}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{C}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \sum_{t=1}^{T} \mathbf{H}_{Nt}(\boldsymbol{\theta}), \tag{12}$$

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

$$\mathbf{B}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \sum_{t=1}^{T} \mathbf{B}_{Nt}(\boldsymbol{\theta}), \tag{13}$$

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 recalling $\boldsymbol{\xi}_t \equiv \mathbf{Y}_t - \boldsymbol{\lambda}_t$. Expectation is taken with respect to the stationary distribution of $\{\mathbf{Y}_t\}$. Moreover, the theoretical counterpart of the Hessian and information matrices respectively, are the following.

$$\mathbf{H}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right], \tag{14}$$

$$\mathbf{B}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right]. \tag{15}$$

Similarly for the log-linear PNAR model, we have that the score function is given by:

$$\mathbf{S}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{i=1}^{N} \left(Y_{i,t} - \exp(\nu_{i,t}(\boldsymbol{\theta})) \right) \frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\mathbf{Y}_t - \exp(\boldsymbol{\nu}_t(\boldsymbol{\theta})) \right), \tag{16}$$

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

$$\mathbf{H}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}, \tag{17}$$

$$\mathbf{B}_{NT}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T},$$

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 $i = 1, \dots, N$ 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,

$$\mathbf{H}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right], \tag{18}$$

$$\mathbf{B}_N(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right], \tag{19}$$

are respectively (minus) the Hessian matrix and the information matrix.

3.1. Linear Model Inference

Recall (10). We drop the dependence on $\boldsymbol{\theta}$ when a quantity is evaluated at $\boldsymbol{\theta}_0$. For ease of presentation, consider model (2) with first moment $\mathrm{E}(\mathbf{Y}_t) = \boldsymbol{\mu} = \mu \mathbf{1}_N$ where $\mu = \beta_0/(1 - \beta_1 - \beta_2)$ (see Section S-1 in Supplement SM). Moreover, the elementwise absolute value of the error covariance matrix is defined as $\boldsymbol{\Sigma\xi} = \mathrm{E}\left|\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T\right|_v$. Define the following expectations $\Pi_{222} = N^{-1}\sum_{i=1}^N \mathrm{E}[(\mathbf{w}_i^T(\mathbf{Y}_{t-1} - \boldsymbol{\mu}))^3/\lambda_{i,t}]$, $\Pi_{223} = N^{-1}\sum_{i=1}^N \mathrm{E}[(\mathbf{w}_i^T(\mathbf{Y}_{t-1} - \boldsymbol{\mu}))^2 Y_{i,t-1}/\lambda_{i,t}]$, $\Pi_{233} = N^{-1}\sum_{i=1}^N \mathrm{E}[\mathbf{w}_i^T(\mathbf{Y}_{t-1} - \boldsymbol{\mu}) Y_{i,t-1}^2/\lambda_{i,t}]$, $\Pi_{333} = N^{-1}\sum_{i=1}^N \mathrm{E}[Y_{i,t-1}^3/\lambda_{i,t}]$. Those expectations constitute summands for some of the elements of the expected third derivative matrix of $l_{i,t}(\boldsymbol{\theta})$. Consider the set $\Omega_d = \{(2,2,2), (2,2,3), (2,3,3), (3,3,3)\}$, $(j^*, l^*, k^*) = \arg\max_{1 \le j,l,k \le m} \left|N^{-1}\sum_{i=1}^N \partial^3 l_{i,t}(\boldsymbol{\theta})/\partial\theta_j \partial\theta_l \partial\theta_k\right|$ is the set of indices where the absolute value of the third derivative is maximum. Assume the following:

B1 The process $\left\{\boldsymbol{\xi}_t,\ \mathcal{F}_t^N :\ N \in \mathbb{N}, t \in \mathbb{Z}\right\}$ is α-mixing, where $\mathcal{F}_t^N = \sigma\left(\xi_{i,s} :\ 1 \le i \le N, s \le t\right)$.

B2 Let $\mathbf{W}$ be a sequence of matrices with non-random entries indexed by N.

- B2.1 Consider $\mathbf{W}$ as a transition probability matrix of a Markov chain, whose state space is defined as the set of all the nodes in the network (i.e., $\{1, \dots, N\}$). The Markov chain is assumed to be irreducible and aperiodic. Further, define $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^T \in \mathbb{R}^N$ as the stationary distribution of the Markov chain, where $\pi_i \ge 0$, $\sum_{i=1}^N \pi_i = 1$ and $\boldsymbol{\pi} = \mathbf{W}^T\boldsymbol{\pi}$. Furthermore, assume that $\lambda_{\max}(\boldsymbol{\Sigma\xi})\sum_{i=1}^N \pi_i^2 \to 0$ as $N \to \infty$.
- B2.2 Define $\mathbf{W}^* = \mathbf{W} + \mathbf{W}^T$ and assume that $\lambda_{\max}(\mathbf{W}^*) = \mathcal{O}(\log N)$ and $\lambda_{\max}(\boldsymbol{\Sigma\xi}) = \mathcal{O}((\log N)^\delta)$, for some $\delta \ge 1$.

B3 Set $\boldsymbol{\Lambda} = \mathrm{E}(\mathbf{D}_t^{-1})$, $\overline{\boldsymbol{\Gamma}}(0) = \mathrm{E}[\mathbf{D}_t^{-1/2}(\mathbf{Y}_{t-1} - \boldsymbol{\mu})(\mathbf{Y}_{t-1} - \boldsymbol{\mu})^T\mathbf{D}_t^{-1/2}]$ and $\boldsymbol{\Delta}(0) = \mathrm{E}[\mathbf{D}_t^{-1/2}\mathbf{W}(\mathbf{Y}_{t-1} - \boldsymbol{\mu})(\mathbf{Y}_{t-1} - \boldsymbol{\mu})^T\mathbf{W}^T\mathbf{D}_t^{-1/2}]$. Assume the following limits exist: $d_1 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\,(\boldsymbol{\Lambda})$, $d_2 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\overline{\boldsymbol{\Gamma}}(0)\right]$, $d_3 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\mathbf{W}\overline{\boldsymbol{\Gamma}}(0)\right]$, $d_4 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\,[\boldsymbol{\Delta}(0)]$ and, if $(j^*, l^*, k^*) \in \Omega_d$, $d_* = \lim_{N\to\infty} \Pi_{j^*,l^*,k^*}$.

Assumption B1 (see Doukhan, 1994) is a mixing condition. Recall that $\boldsymbol{\xi}_t$ is an α-mixing array if, namely,

$$\alpha(J) = \sup_{N\in\mathbb{N}} \alpha_N(J) = \sup_{t\in\mathbb{Z}, N\in\mathbb{N}} \sup_{A\in\mathcal{F}_{-\infty,t}^N, B\in\mathcal{F}_{t+J,\infty}^N} |\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| \xrightarrow{J\to\infty} 0,$$

where $\mathcal{F}_t^N \equiv \mathcal{F}_{-\infty,t}^N = \sigma\left(\xi_{i,s} :\ 1 \le i \le N, s \le t\right)$, $\mathcal{F}_{t+J,\infty}^N = \sigma\left(\xi_{i,s} :\ 1 \le i \le N, s \ge t + J\right)$. This assumption holds true for the simple example of $\boldsymbol{\xi}_t \sim IID(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\xi}_t$ is constructed by the copula method proposed in this article. In this case, the noise is independent over time but it is non contemporaneous independent. Another example

DOI: 10.1111/jtsa.12728

would be all the processes which satisfy $\alpha_N(J) \leq f(J)$, where $f(J)$ is some function which does not depend on N, such that $f(J) \to 0$ as $J \to \infty$.

Assumption B2 on the network structure implies that the edges between nodes are known and as N increases to $N+1$ an additional node is added with some edges to the previous N nodes, but the edges among the previous N nodes do not change. Moreover, it requires some uniformity conditions, and it is equivalent set of conditions as Zhu *et al.* (2019, C2, C2.1-C2.2). Finally, B2.2 requires that the network structure admits certain uniformity property ($\lambda_{\max}(\mathbf{W}^*)$ diverges slowly). Zhu *et al.* (2017, supp. mat., sec. 7.1-7.3) found empirically that this is the case for several network models. In our case, regularity assumptions on the structure of dependence among the errors, when the network grows, are required by imposing that the diverging rate of $\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}})$ will be slower than order $\mathcal{O}(N)$, in B2.2, and its product with the squared sum of the stationary distribution of the chain, $\boldsymbol{\pi}$, will tend to 0, in B2.1. We give an empirical verification of such conditions in Section S-4 of the Supplement SM. In the continuous-valued case introduced in Zhu *et al.* (2017) such assumptions are not necessary because the errors are i.i.d with common variance σ^2. Moreover, in this case, the absolute value is no more required because $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \mathrm{E}\left(\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T\right) = \sigma^2 \mathbf{I}_N$.

The conditions outlined in B3 are law of large numbers-like assumptions, which are quite standard in the existing literature, since little is known about the behavior of the process as $N \to \infty$. These assumptions are required to guarantee that the Hessian matrix (12) converges to a matrix which exists. Section S-4 in Supplement SM includes numerical study examples showing the validity of these limits. If OLS estimation with i.i.d errors was performed, conditions B3 would correspond exactly to those in Zhu *et al.* (2017, C3).

Lemma 1. *For the linear model (2), suppose $\beta_1 + \beta_2 < 1$ and B1–B3 hold. Consider $\mathbf{S}_{NT}$ and $\mathbf{H}_{NT}$ defined as in (11) and (12) respectively. Then, as $\{N, T_N\} \to \infty$*

1. $(NT_N)^{-1}\mathbf{H}_{NT_N} \xrightarrow{p} \mathbf{H}$,
2. $(NT_N)^{-1}\mathbf{S}_{NT_N} \xrightarrow{p} \mathbf{0}_m$,
3. $\max_{j,l,k} \sup_{\boldsymbol{\theta} \in \mathcal{O}(\boldsymbol{\theta}_0)} \left| (NT_N)^{-1} \sum_{t=1}^{T_N} \sum_{i=1}^{N} \frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l \partial \theta_k} \right| \leq M_{NT_N} \xrightarrow{p} M$,

where $\mathcal{O}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_2 < \delta\}$ is a neighborhood of $\boldsymbol{\theta}_0$, $M_{NT_N} := (NT_N)^{-1} \sum_{t=1}^{T_N} \sum_{i=1}^{N} m_{i,t}$, M is a finite constant, $\mathbf{H} = \lim_{N\to\infty} N^{-1}\mathbf{H}_N$ is non singular and

$$\mathbf{H} = \begin{pmatrix} d_1 & \mu d_1 & \mu d_1 \\ & \mu^2 d_1 + d_4 & \mu^2 d_1 + d_3 \\ & & \mu^2 d_1 + d_2 \end{pmatrix}. \tag{20}$$

Some preliminary results required to show the lemma are proved in Section S-3.1 in Supplement SM. The proof of Lemma 1 is given in Appendix A.2.

Consider now the following conditions:

B3′ Set $\boldsymbol{\Lambda}_t = \boldsymbol{\Sigma}_t^{1/2}\mathbf{D}_t^{-1}$, $\boldsymbol{\Lambda} = \mathrm{E}(\boldsymbol{\Lambda}_t^T \boldsymbol{\Lambda}_t)$, $\overline{\boldsymbol{\Gamma}}(0) = \mathrm{E}[\boldsymbol{\Lambda}_t(\mathbf{Y}_{t-1} - \boldsymbol{\mu})(\mathbf{Y}_{t-1} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}_t^T]$ and $\boldsymbol{\Delta}(0) = \mathrm{E}[\boldsymbol{\Lambda}_t \mathbf{W}(\mathbf{Y}_{t-1} - \boldsymbol{\mu})(\mathbf{Y}_{t-1} - \boldsymbol{\mu})^T \mathbf{W}^T \boldsymbol{\Lambda}_t^T]$. Assume that the following limits exist:
$f_1 = \lim_{N\to\infty} N^{-1}\left(\mathbf{1}_N^T \boldsymbol{\Lambda} \mathbf{1}_N\right), f_2 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\overline{\boldsymbol{\Gamma}}(0)\right], f_3 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\mathbf{W}\overline{\boldsymbol{\Gamma}}(0)\right]$,
$f_4 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\boldsymbol{\Delta}(0)\right]$ and, if $(j^*, l^*, k^*) \in \Omega_d$, $d_* = \lim_{N\to\infty} \Pi_{j^*,l^*,k^*}$.

B4 There exists a non-negative, non-increasing sequence $\{\varphi_h\}_{h=1,\ldots,\infty}$ such that $\sum_{h=1}^{\infty} \varphi_h = \Phi < \infty$ and, for $i < j$, almost surely

$$\left|\mathrm{Corr}(Y_{i,t}, Y_{j,t} \mid \mathcal{F}_{t-1})\right| \leq \varphi_{j-i}, \tag{21}$$

Condition B3′ is simply an extension of assumption B3, required for the convergence of the conditional information matrix (13) to a valid limiting information matrix, see (22) below. More precisely, the reader can verify

that B3 is just a special case of B3′, when $\mathbf{\Sigma}_t = \mathbf{D}_t$. The main reason that this assumption is introduced is that, for the QMLE, the conditional information matrix and the Hessian matrix are, in general, different. This does not occur in the case studied by Zhu *et al.* (2017). Analogously to B3, when $\mathbf{Y}_t$ is continuous-valued random vector, and we deal with IID errors $\boldsymbol{\xi}_t$, B3′ reduces again to the conditions in Zhu *et al.* (2017, C3).

Assumption B4 could be considered as a contemporaneous weak dependence condition. Indeed, even in the very simple case of independence model, i.e. $\lambda_{i,t} = \beta_0$, for all $i = 1, \dots, N$, the reader can easily verify that, without any further constraints, $N^{-1}\mathbf{B}_N = \mathcal{O}(N)$, so the limiting variance of the estimator will eventually diverge, since it depends on the limit of the conditional information matrix. Instead, under B4, $N^{-1}\mathbf{B}_N = \mathcal{O}(1)$, and the existence of the limiting covariance matrix can be shown, as in Lemma 2 and Theorem 3 below. Insights about weak dependence conditions have been stated in Zhu *et al.* (2017, p. 1102). When the errors are independent over different nodes and the past (Zhu *et al.*, 2017, C1), B4 is trivially satisfied, since $\left|\mathrm{Cov}(Y_{i,t}, Y_{j,t} \mid \mathcal{F}_{t-1})\right| = \left|\mathrm{E}(\xi_{i,t}\xi_{j,t})\right| = 0$, for $i \neq j$. See Section S-5 of the Supplement SM, for an example where B4 is empirically verified. Define $\boldsymbol{\eta} \in \mathbb{R}^m$, a non-null real-valued vector.

Lemma 2. *For the linear model (2), suppose $\beta_1 + \beta_2 < 1$ and B1-B2, B3′-B4 hold. Consider $\mathbf{S}_{NT}$ and $\mathbf{B}_{NT}$ defined as in (11) and (13) respectively. Assume $N^{-2}\mathrm{E}(\boldsymbol{\eta}^T\mathbf{s}_{Nt})^4 < \infty$. Then, as $\{N, T_N\} \to \infty$*

1. $(NT_N)^{-1}\mathbf{B}_{NT_N} \xrightarrow{p} \mathbf{B}$,
2. $(NT_N)^{-\frac{1}{2}}\mathbf{S}_{NT_N} \xrightarrow{d} N(\mathbf{0}_m, \mathbf{B})$,

where $\mathbf{B} = \lim_{N\to\infty} N^{-1}\mathbf{B}_N$ and

$$\mathbf{B} = \begin{pmatrix} f_1 & \mu f_1 & \mu f_1 \\ & \mu^2 f_1 + f_4 & \mu^2 f_1 + f_3 \\ & & \mu^2 f_1 + f_2 \end{pmatrix}. \tag{22}$$

Note that the assumption $N^{-2}\mathrm{E}(\boldsymbol{\eta}^T\mathbf{s}_{Nt})^4 < \infty$ is not implied by condition B4 which is satisfied provided that (21) holds true for higher-order moments of the vectors $\{\mathbf{Y}_t\}$; See Section S-6 in Supplement SM more.

Theorem 3. *Consider model (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 the conditions of Lemma 1 and 2 hold. Then, there exists a fixed open neighborhood $\mathcal{O}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_2 < \delta\}$ of $\boldsymbol{\theta}_0$ such that with probability tending to 1 as $\{N, T_N\} \to \infty$, for the score function (11), the equation $S_{NT_N}(\boldsymbol{\theta}) = \mathbf{0}_m$ has a unique solution, called $\hat{\boldsymbol{\theta}}$, which is consistent and asymptotically normal*:

$$\sqrt{NT_N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(\mathbf{0}_m, \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}\right).$$

The extension of Theorem 3 to the general order linear PNAR(p) model is immediate, by using the well-known VAR(1) companion matrix; see Section (S-1) in Supplement SM. Assumptions B1, B2 and B4 remain substantially unaffected, using Debaly and Truquet (2021, lemma 1.1). B3-B3′ can be suitably rearranged similarly to Zhu *et al.* (2017, C4) and the result follows by Zhu *et al.* (2017, supp. mat., sec. 4). We omit the details.

Remark 5. A standard asymptotic inference result, with $T \to \infty$, is obtained for the QMLE $\hat{\boldsymbol{\theta}}$, where the 'sandwich' covariance is $\mathbf{H}_N^{-1}\mathbf{B}_N\mathbf{H}_N^{-1}$, by Theorem 3, as a special case, when N is fixed. This result requires only the stationarity conditions of Proposition 1, the compactness of the parameter space, and assuming that the true value of the parameters belongs to its interior. Such result is proved along the lines of theorem 4.1 in Fokianos *et al.* (2020). Similar comments apply also for the log-linear model below. The case, where T fixed and N diverging, cannot be studied in the framework we consider, since the convergence of the quantities involved in Lemmas 1 and 2 requires both indexes to diverge together. For details see also the related proofs in the Appendix A.2. This is empirically confirmed by some numerical bias found in the simulations of Section 4.1, when T is small compared to N.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

Remark 6. It is worth pointing out that model (2) may be extended by including a feedback process such as as

$$\mathbf{Y}_t|\mathcal{F}_{t-1} \sim MCP(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \boldsymbol{\beta}_0 + \mathbf{G}\mathbf{Y}_{t-1} + \mathbf{J}\boldsymbol{\lambda}_{t-1}, \tag{23}$$

where $\mathbf{J} = \alpha_1\mathbf{W} + \alpha_2\mathbf{I}_N$ and $\alpha_1, \alpha_2 \geq 0$ will be a network and autoregressive coefficients respectively, for the past values of the conditional mean process. Such extension is suitable when the mean process $\boldsymbol{\lambda}_t$ depends on the whole past history of the count process. When the network dimension is fixed, model (23) is just a special case of Fokianos *et al.* (2020, eq. 3), with a specific neighbor structure of the coefficients matrices therein. The stability conditions and asymptotic properties of the QMLE follow immediately. Note that (23) implies $\boldsymbol{\lambda}_t = f(\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \dots,)$, so all likelihood based quantities are evaluated recursively (for more, see Fokianos *et al.*, 2020, eq. 12), Therefore, when $N \to \infty$, verification of Assumptions like B1–B4, which guarantee good large-sample properties of the corresponding estimators, is quite challenging problem because the dimension of the hidden process grows. See Section S.3.1 in Supplement SM for comparison. Similarly, the log-linear model (6) can be extended by including the process $\boldsymbol{\nu}_{t-1}$ in the right-hand side but the same problem persists.

3.2. Log-Linear Model Inference

We now state the analogous result for the log-linear model (7) and the notation corresponds to equations (16)–(19). Set $\mathbf{Z}_t = \log(\mathbf{1}_N + \mathbf{Y}_t)$ and recall that $\mathrm{E}(\mathbf{Z}_t) \approx \boldsymbol{\mu}$ by the discussion below equation (7). Define $\sigma_{ij} = \mathrm{E}(\xi_{i,t}\xi_{j,t})$ the single element of the error covariance matrix, and $\Pi^L_{222} = N^{-1}\sum_{i=1}^N \mathrm{E}[\exp(\nu_{i,t})(\mathbf{w}_i^T(\mathbf{Z}_{t-1} - \boldsymbol{\mu}))^3]$, $\Pi^L_{223} = N^{-1}\sum_{i=1}^N \mathrm{E}[\exp(\nu_{i,t})(\mathbf{w}_i^T(\mathbf{Z}_{t-1} - \boldsymbol{\mu}))^2 Y_{i,t-1}]$, $\Pi^L_{233} = N^{-1}\sum_{i=1}^N \mathrm{E}[\exp(\nu_{i,t})\mathbf{w}_i^T(\mathbf{Y}_{t-1} - \boldsymbol{\mu})Y^2_{i,t-1}]$, $\Pi^L_{333} = N^{-1}\sum_{i=1}^N \mathrm{E}[\exp(\nu_{i,t})Y^3_{i,t-1}]$. Assumption B1^L is the same as assumption B1 in the linear model. This holds also for B2^L, by considering $\boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \mathrm{E}\left|\boldsymbol{\psi}_t\boldsymbol{\psi}_t^T\right|_v$ instead of $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ in B2 above.

B3^L Set $\overline{\boldsymbol{\Gamma}}^L(0) = \mathrm{E}[\boldsymbol{\Sigma}_t^{1/2}(\mathbf{Z}_{t-1} - \boldsymbol{\mu})(\mathbf{Z}_{t-1} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}_t^{1/2}]$ and $\boldsymbol{\Delta}^L(0) = \mathrm{E}[\boldsymbol{\Sigma}_t^{1/2}\mathbf{W}(\mathbf{Z}_{t-1} - \boldsymbol{\mu})(\mathbf{Z}_{t-1} - \boldsymbol{\mu})^T\mathbf{W}^T\boldsymbol{\Sigma}_t^{1/2}]$. Assume the following limits exist: $l_1 = \lim_{N\to\infty} N^{-1}\mathrm{E}[\mathbf{1}_N^T\mathbf{D}_t\mathbf{W}(\mathbf{Z}_{t-1} - \boldsymbol{\mu})]$, $l_2 = \lim_{N\to\infty} N^{-1}\mathrm{E}[\mathbf{1}_N^T\mathbf{D}_t(\mathbf{Z}_{t-1} - \boldsymbol{\mu})]$, $\varsigma = \lim_{N\to\infty} N^{-1}\sum_{i\neq j}\sigma_{ij}$, $g_3 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\overline{\boldsymbol{\Gamma}}^L(0)\right]$, $g_4 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\mathbf{W}\overline{\boldsymbol{\Gamma}}^L(0)\right]$, $g_5 = \lim_{N\to\infty} N^{-1}\mathrm{tr}\left[\boldsymbol{\Delta}^L(0)\right]$ and, if $(j^*, l^*, k^*) \in \Omega_d$, $d_* = \lim_{N\to\infty}\Pi^L_{j^*,l^*,k^*}$.

B4^L There exists a non-negative, non-increasing sequence $\{\phi_h\}_{h=1,\dots,\infty}$ such that $\sum_{h=1}^\infty \phi_h = \Phi < \infty$ and, for $i < j$, almost surely

$$\left|\mathrm{Cov}(Y_{i,t}, Y_{j,t} \mid \mathcal{F}_{t-1})\right| \leq \phi_{j-i}. \tag{24}$$

The same remarks made for the case of linear model hold true in this case as well. Condition B4^L has been stated in terms of conditional covariances instead of correlations. This is simply due to the different form of the information matrix (19), which only includes the conditional covariance matrix $\boldsymbol{\Sigma}_t$. In contrast the linear model information matrix which corresponds to (15) is given by $\mathbf{B}_N = \mathrm{E}(\partial\boldsymbol{\lambda}_t^T/\partial\boldsymbol{\theta}\mathbf{D}_t^{-1/2}\mathbf{R}_t\mathbf{D}_t^{-1/2}\partial\boldsymbol{\lambda}_t/\partial\boldsymbol{\theta}^T)$, where $\mathbf{R}_t = \mathbf{D}_t^{-1/2}\boldsymbol{\Sigma}_t\mathbf{D}_t^{-1/2}$ is conditional correlation matrix, and $\mathbf{D}_t^{-1/2} \prec \beta_0^{-1}\mathbf{I}_N$ (elementwise), so that working with the correlations is more natural and convenient. Numerical verification of assumptions B2^L-B3^L is given in Section S-4 in Supplement SM, and complement the results of the linear model. Recall that $\boldsymbol{\eta} \in \mathbb{R}^m$, denotes a non-null real-valued vector.

Lemma 3. For the log-linear model (7), suppose $|\beta_1| + |\beta_2| < 1$ and B1^L-B4^L hold. Consider $\mathbf{S}_{NT}$ and $\mathbf{H}_{NT}$ defined as in (16) and (17) respectively. Assume $N^{-2}\mathrm{E}(\boldsymbol{\eta}^T\mathbf{s}_{Nt})^4 < \infty$. Then, as $\{N, T_N\} \to \infty$

1. $(NT_N)^{-1}\mathbf{H}_{NT_N} \xrightarrow{p} \mathbf{H}$,
2. $(NT_N)^{-\frac{1}{2}}\mathbf{S}_{NT_N} \xrightarrow{d} N(\mathbf{0}_m, \mathbf{B})$,

3. $\max_{j,l,k} \sup_{\theta \in \mathcal{O}(\theta_0)} \left| (NT_N)^{-1} \sum_{t=1}^{T_N} \sum_{i=1}^{N} \frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l \partial \theta_k} \right| \leq M_{NT_N} \xrightarrow{p} M$,

where $\mathbf{H} = \lim_{N\to\infty} N^{-1}\mathbf{H}_N$ is non-singular and

$$\mathbf{H} = \begin{pmatrix} \mu_y & l_1^* & l_2^* \\ & \mu(l_1^* + l_1) + l_5 & \mu(l_1^* + l_2) + l_4 \\ & & \mu(l_2^* + l_2) + l_3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mu_y^* & g_1^* & g_2^* \\ & \mu(g_1^* + l_1) + g_5 & \mu(g_1^* + l_2) + g_4 \\ & & \mu(g_2^* + l_2) + g_3 \end{pmatrix}, \tag{25}$$

where $\mu_y = \mathrm{E}(Y_{i,t})$, $l_1^* = \mu\mu_y + l_1$, $l_2^* = \mu\mu_y + l_2$, (l_3, l_4, l_5) equal (g_3, g_4, g_5) respectively, when $\boldsymbol{\Sigma}_t = \mathbf{D}_t$, $\mu_y^* = \mu_y + \varsigma$, $g_1^* = \mu\mu_y^* + l_1$ and $g_2^* = \mu\mu_y^* + l_2$.

Theorem 4. *Consider model (7). 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 3 hold. Then, there exists a fixed open neighborhood $\mathcal{O}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_2 < \delta\}$ of $\boldsymbol{\theta}_0$ such that with probability tending to 1 as $\{N, T_N\} \to \infty$, for the score function (16), the equation $S_{NT_N}(\boldsymbol{\theta}) = \mathbf{0}_m$ has a unique solution, called $\hat{\boldsymbol{\theta}}$, which is consistent and asymptotically normal*:

$$\sqrt{NT_N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}_m, \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}).$$

The conclusion follows arguing as in the proof of Theorem 3 above. An analogous result can be established for $p > 1$, since also $\log(\mathbf{1}_N + \mathbf{Y}_t)$ in (7) can be approximately rewritten as a VAR(1) model; see also Section S-1 in Supplement SM.

3.3. Estimation of Covariance Matrix

We provide a consistent estimator for the limiting covariance matrix of the QMLE. Toward this goal, define the following matrix

$$\hat{\mathbf{B}}_{NT}(\hat{\boldsymbol{\theta}}) = \sum_{t=1}^{T} \mathbf{s}_{Nt}(\hat{\boldsymbol{\theta}}) \mathbf{s}_{Nt}^T(\hat{\boldsymbol{\theta}}). \tag{26}$$

Let $\mathbf{V} := \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ and $\mathbf{V}_{NT}(\hat{\boldsymbol{\theta}}) := (NT)\mathbf{H}_{NT}^{-1}(\hat{\boldsymbol{\theta}})\hat{\mathbf{B}}_{NT}(\hat{\boldsymbol{\theta}})\mathbf{H}_{NT}^{-1}(\hat{\boldsymbol{\theta}})$. The following theorem shows how to consistently estimate the covariance matrix obtained by Theorems 3 and 4 by using the usual sandwich estimator.

Theorem 5. Consider model (2) (respectively, model (7)). Suppose the conditions of Theorem 3 (respectively, Theorem 4) hold true. Then, as $\{N, T_N\} \to \infty$, $\mathbf{V}_{NT_N}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{V}$.

3.4. Effect of Network Misspecification

We now study the effect of network misspecification. Consider, for instance, model (2). Suppose the data $Y_{i,t}$ are generated by the true adjacency matrix $\mathbf{A}$. From Section 3, $\hat{\boldsymbol{\theta}}$ is consistent estimator of $\boldsymbol{\theta}$. Suppose that the adjacency matrix $\mathbf{A}$ is misspecified and the true network matrix is $\mathbf{A}^* = (a_{ij}^*)$. Accordingly, let $\mathbf{W}^* = (w_{ij}^*)$ be the row-normalized adjacency matrix $\mathbf{A}^*$ and $\lambda_{i,t}^*(\boldsymbol{\theta})$ defined as in (1) but with the elements of $\mathbf{W}^*$ instead of $\mathbf{W}$. Then, the QMLE, in this case, is given by $\hat{\boldsymbol{\theta}}^* = \arg\max_{\theta\in\Lambda} l_{NT}^*(\boldsymbol{\theta})$ where $l_{NT}^*(\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\left(Y_{i,t}\log\lambda_{i,t}^*(\boldsymbol{\theta}) - \lambda_{i,t}^*(\boldsymbol{\theta})\right)$.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

Corollary 1. *Assume the conditions of Theorem 3 hold. Define* $\Delta_N(\mathbf{W}, \mathbf{W}^*) = \sum_{i,j=1}^{N} \left| w_{ij} - w_{ij}^* \right|$ *the total amount of misspecification of* $\mathbf{W}$. *Assume* $\Delta_N(\mathbf{W}, \mathbf{W}^*) = o(1)$, *then as* $\{N, T_N\} \to \infty$ $\hat{\theta}^* \xrightarrow{p} \theta_0$.

The proof is postponed to Section S-3.5 of Supplement SM. Corollary 1 shows that, under network misspecification, the QMLE is still consistent estimator provided that the amount of misspecification is under control, i.e., $\lim_{N\to\infty} \Delta_N(\mathbf{W}, \mathbf{W}^*) = 0$. For example $\Delta_N(\mathbf{W}, \mathbf{W}^*) \leq C/\sqrt{N}$ or $C/\log(N)$ for some constant $C > 0$ implies Corollary 1. Similar results hold for $p > 1$ and log-linear models (9).

3.5. Further Discussion

Some further issues are described next.

3.5.1. Copula Estimation

Copula estimation is briefly discussed in the Supplement SM but a thorough study of the problem requires separate treatment. In particular, Section S-9 of the Supplement SM contains results of a simulation study after employing a heuristic parametric bootstrap estimation algorithm. Such method potentially can be useful to select an adequate copula structure and provide an estimator of the associated copula parameter.

3.5.2. Efficiency of QMLE

The QMLE based on (10), is general inefficient. Therefore, in Section S-7 of the Supplement SM, a novel regression estimator is proposed by considering a two-step Generalized Estimating Equations (GEE). During the first step, the mean parameters are estimated by QMLE and employed to compute a working weighting covariance matrix. A second step of estimation is then carried out by employing the obtained weighting matrix. Numerical studies show that the GEE is more efficient than the QMLE, especially when there exists considerable correlation among the counts. In a recent paper by Aknouche and Francq (2023) a similar kind of estimators, but for univariate models, have been shown to be optimal QMLEs, under suitable regularity condition.

3.5.3. State-space modeling

An alternative approach to the methodology developed in this article is to consider a state-space model as in the work by Zhang *et al.* (2017), for example. These authors develop methodology for the log-linear model (6) by adding Gaussian noise to the right hand side of the model defining equation. In addition, marginal counts are assumed to be Poisson($\lambda_{i,t}$) distributed – recall the notation of Section 2. The authors develop particle filtering and smoothing methods together with Monte Carlo Expectation Maximization algorithm to advance inference. A fully Bayesian approach, related to network models, is taken by Chen *et al.* (2019) who introduce models within the framework of dynamic GLM (see West and Harrison, 1997), that include time-varying covariates for Poisson conditionally distributed time series; for more on the Bayesian point of view see West (2020).

4. APPLICATIONS

4.1. Simulations

We study the finite sample behavior of the QMLE for models (4) and (9). We run a simulation study with $S = 1000$ repetitions and different time series length and network dimension. We consider the cases $p = (1, 2)$. The adjacency matrix is generated by using one of the most popular network structure, the stochastic block model (SBM):

Example 1. (SBM). A block label ($k = 1, \ldots, K = 5$) is assigned for each node with equal probability and K is the total number of blocks. Then, set $\mathrm{P}(a_{ij} = 1) = N^{-0.3}$ if i and j belong to the same block, and $\mathrm{P}(a_{ij} = 1) = N^{-1}$ otherwise. Practically, the model assumes that nodes within the same block are more likely to be connected with respect to nodes from different blocks.

DOI: 10.1111/jtsa.12728 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

For details on SBM see Wang and Wong (1987), Nowicki and Snijders (2001) and Zhao *et al.* (2012), among others. The SBM model with $K = 5$ blocks is generated by using the `igraph` `R` package (Csardi and Nepusz, 2006). The network density is set equal to 1%. We performed simulations with a network density equal to 0.3% and 0.5%, as well, but we obtained similar results, hence we do not report them here. The parameters are set to $(\beta_0, \beta_1, \beta_2)^T = (0.2, 0.3, 0.2)^T$. The observed time series are generated using the copula-based algorithm of Section 2.1. The specified copula is Gaussian, say $C^{\text{Ga}}_{\mathbf{R}}(\dots)$, with correlation matrix $\mathbf{R} = (R_{ij})$, where $R_{ij} = \rho^{|i-j|}$, the so called first-order autoregressive correlation matrix, henceforth AR-1. Then $C^{\text{Ga}}_{\mathbf{R}}(\dots) = C^{\text{Ga}}(\dots, \rho)$. Tables I and II summarize the simulation results for models (2) and (7) respectively. For each simulated dataset, the QMLE estimation of unknown parameters has been computed by using the `R` package `nloptr` (Johnson, 2023). It allows to run constrained optimization; for the linear model (4), for example, the quasi log-likelihood (10) is maximized under the positive parameters constraint. Additional findings are given in Section S-8 of the Supplement SM – Tables S1–S4.

Then, the estimates for parameters and their SEs (in brackets) are obtained by averaging out the results from all simulations; see the first two rows of Tables I and II. The third row below each coefficient shows the percentage frequency of t-tests rejecting $H_0 : \beta = 0$ at nominal level 5% and it is calculated 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 the estimates are close to the real values and the SEs are small for all the cases considered. When there is a strong correlation between count variables $Y_{i,t}$ – see Table I – and T is small when compared to the network size N, then the estimates of the network effect $\hat{\beta}_1$ have slight bias. The same conclusion is drawn from Table S1. Instead, when both T and N are reasonably large (or at least T is large), then the approximation to the true values of the parameters is adequate. This fact confirms the related asymptotic results obtained in Section 3

Table I. Estimators obtained from $S = 1000$ simulations of model (2), for various values of N and T

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	0.296	0.199	0.197	0.292	0.196	0.009	0.007	94.1.0	99.5	95.1
		(0.019)	(0.036)	(0.028)	(0.021)	(0.037)	(0.029)	(0.031)	(0.023)			
		100	100	100	100	100	100	1.4	1.5			
	200	0.200	0.297	0.199	0.197	0.294	0.197	0.008	0.005	93.9	99.9	95.2
		(0.013)	(0.027)	(0.020)	(0.014)	(0.028)	(0.021)	(0.023)	(0.016)			
		100	100	100	100	100	100	1.5	1.6			
100	20	0.203	0.292	0.198	0.196	0.286	0.195	0.015	0.008	93.1	97.1	93.5
		(0.024)	(0.048)	(0.028)	(0.029)	(0.050)	(0.029)	(0.046)	(0.024)			
		100	100	100	100	100	100	2.9	2.2			
	50	0.202	0.294	0.199	0.197	0.290	0.197	0.011	0.005	91.4	98.8	94.1
		(0.015)	(0.032)	(0.018)	(0.018)	(0.033)	(0.019)	(0.031)	(0.015)			
		100	100	100	100	100	100	3.3	2.0			
	100	0.201	0.299	0.200	0.198	0.296	0.198	0.008	0.004	91.9	99.2	94.9
		(0.011)	(0.023)	(0.013)	(0.013)	(0.023)	(0.013)	(0.022)	(0.011)			
		100	100	100	100	100	100	2.0	1.8			
	200	0.200	0.299	0.200	0.198	0.298	0.199	0.005	0.003	92.3	99.7	95.2
		(0.008)	(0.016)	(0.009)	(0.009)	(0.017)	(0.009)	(0.015)	(0.008)			
		100	100	100	100	100	100	2.0	1.6			

Note: Network generated by Ex. 1. Data are generated by using the Gaussian AR-1 copula, with $\rho = 0.5$ and $p = 1$. Model (4) is also fitted using $p = 2$ to check the performance of various information criteria.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

Table II. Estimators obtained from $S = 1000$ simulations of model (7), for various values of N and T.

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	0.206	0.298	0.196	0.208	0.298	0.196	−0.002	−0.001	81.6	97.5	86.1
		(0.061)	(0.040)	(0.034)	(0.072)	(0.041)	(0.035)	(0.040)	(0.034)			
		91.3	100	100	81.3	100	99.9	2.0	2.7			
	200	0.203	0.298	0.199	0.203	0.298	0.199	0.001	−0.001	80.7	98.9	85.8
		(0.043)	(0.030)	(0.025)	(0.049)	(0.032)	(0.025)	(0.032)	(0.024)			
		99.5	100	100	98.1	100	100	2.3	2.4			
100	20	0.209	0.292	0.196	0.215	0.293	0.197	−0.006	−0.002	74.6	88.2	80.7
		(0.082)	(0.069)	(0.036)	(0.097)	(0.069)	(0.037)	(0.067)	(0.036)			
		70.0	97.5	99.9	59.7	97.4	99.9	3.8	3.3			
	50	0.204	0.296	0.200	0.207	0.296	0.200	−0.004	−0.001	78.4	94.6	86.6
		(0.053)	(0.045)	(0.023)	(0.065)	(0.045)	(0.023)	(0.045)	(0.022)			
		96.3	100	100	86.9	100	100	2.9	2.3			
	100	0.203	0.297	0.199	0.204	0.297	0.200	0.000	−0.001	78.9	97.2	85.7
		(0.037)	(0.031)	(0.016)	(0.046)	(0.032)	(0.016)	(0.031)	(0.016)			
		100	100	100	99.4	100	100	3.1	2.0			
	200	0.201	0.300	0.199	0.203	0.300	0.199	−0.002	0.000	80.5	97.5	88.1
		(0.026)	(0.022)	(0.011)	(0.033)	(0.022)	(0.011)	(0.022)	(0.011)			
		100	100	100	100	100	100	2.9	2.7			

Note: Network generated by Ex. 1. Data are generated by using the Gaussian AR-1 copula, with $\rho = 0.5$ and $p = 1$. Model (9) is also fitted using $p = 2$ to check the performance of various information criteria.

by requiring $N \to \infty$ and $T_N \to \infty$. Standard errors reduce as T increases. Regarding estimators of the log-linear model (see Tables II and S3), we obtain similar results.

The t-tests and percentage of right selections due to various information criteria provide empirical confirmation for the model selection procedure. Based on these results, the BIC provides the best selection procedure for the case of the linear model; its success selection rate is about 99%; this is so because it tends to select models with fewer parameters. In sharp contrast, the AIC is not performing as well as BIC but still selects the right model around 92% of time. The QIC provides a good balance between the other two information criteria; its value is around 95%. Moreover, it has the advantage to be more robust, especially when employed to misspecified models. This fact is further confirmed by the results concerning the log-linear model, even though the rate of right selections for the QIC does not exceed 88%. 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 Tables S2 and S4). Moreover, the slight bias reported, for some coefficients, when $\rho > 0$, is not observed in this case. Intuitively, the reason lies on the complexity of the network relations, which does not grow with N, since variables concerning different nodes are independent. Furthermore, the QMLE for this case coincides to the true likelihood function. From the QQ-plot shown in Figures S13 and S14 we can conclude that, with N and T large enough, the asserted asymptotic normality is quite adequate. A more extensive discussion and further simulation results can be found in Section S-8 of the Supplement SM.

4.2. Data analysis

The application on real data concerns the monthly number of burglaries on the south side of Chicago from 2010 to 2015 ($T = 72$). The counts are registered for the $N = 552$ census block groups. The data are taken by Clark and Dixon (2021), https://github.com/nick3703/Chicago-Data. The undirected network structure arises naturally, as an edge between block i and j is set if the locations share (at least) a border. In this case, the network connection

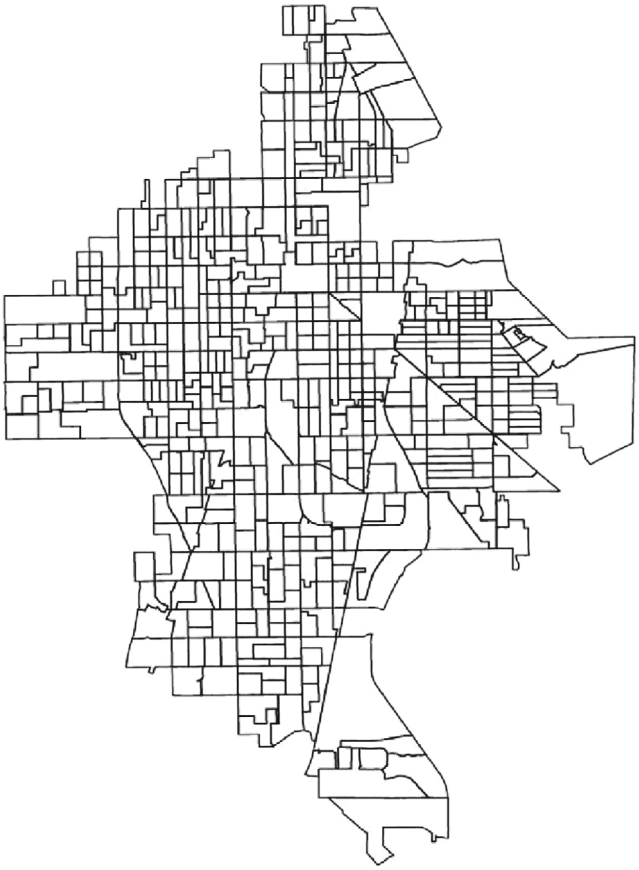

Figure 1. Census block groups in South Chicago

is well-represented by the geographic map of the census blocks in Figure 1. The density of the network is 1.74%. The median degree is 5.

Some time series of burglaries are plotted in Figure 2. The maximum number of burglaries in a month in a census block is 17. We fit the linear and log-linear PNAR(1) and PNAR(2) models. The results are summarized in Tables III and IV. All fitted models produce significant results. 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. The lagged effects have a positive 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 analyzed. All the information criteria select the PNAR(2) models, in accordance with the significance of the estimates.

We compare the out-sample forecasting performance of the linear PNAR model with $p = 1$ vs. a baseline STARMA(1,1) model (Pfeifer and Deutrch, 1980), which after some rearrangement is defined as follows

$$\mathbf{Y}_t = \boldsymbol{\delta}_0 + \left(\phi_1 \mathbf{W} + \phi_0 \mathbf{I}_N\right) \mathbf{Y}_{t-1} + \left(\theta_1 \mathbf{W} + \theta_0 \mathbf{I}_N\right) \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t$ are independent normal vectors, and $\boldsymbol{\delta}_0$, ϕ_i, θ_i, $i = 0, 1$ are unknown parameters. The Root Mean Square Error (RMSE) obtained by both models is computed. For the PNAR model the RMSE is 0.038 which is less than 0.079 obtained by fitting the STARMA(1,1) model. This shows significant accuracy improvement of the prediction for the PNAR(1) model. In addition, PNAR avoids estimation of moving average parameters.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

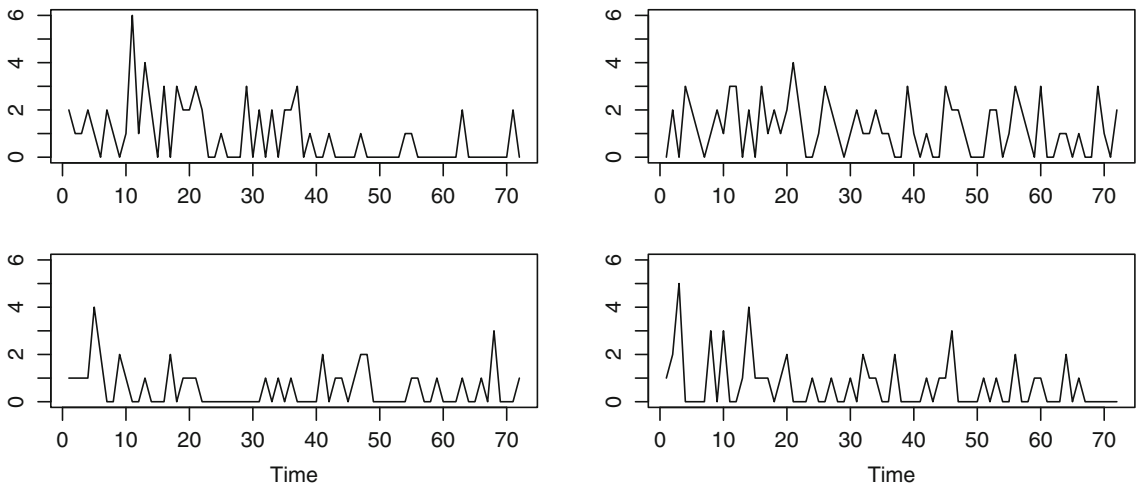

Figure 2. Monthly burglaries count time series for some census block groups

Table III. Estimation results for Chicago crime data

	Estimate	SE ($\times 10^2$)	p-value	Estimate	SE ($\times 10^2$)	p-value
		Linear PNAR(1)			Log-linear PNAR(1)	
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)	
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 IV. Information criteria for Chicago crime data

	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**

Note: Smaller values in bold.

Estimation of the copula is advanced according to the algorithm of Section S-9 of the Supplement SM. The Gaussian AR-1 copula, described in Section 4.1, is compared vs. the Clayton copula, over a grid of values for the associated copula parameter, with 100 bootstrap simulations. As a preliminary step for the estimation of Gaussian AR-1 copula we need to reorder the observations $Y_{i,t}$ for $i = 1, \ldots, N$ to mimic the structure of the AR-1 copula correlation matrix $\mathbf{R} = (R_{ij})$, where $R_{ij} = \rho^{|i-j|}$. A coherent ordering for $Y_{i,t}$ will be the one where the empirical correlation matrix of $\mathbf{Y}_t$, say $\mathbf{R}_e$, contains highest correlations close to the main diagonal and then progressively smaller values where the distance from the main diagonal increases. This is a combinatorial problem and for small N it is not hard to solve it by trying all the possible orderings. However, when N grows, we can recover

DOI: 10.1111/jtsa.12728 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

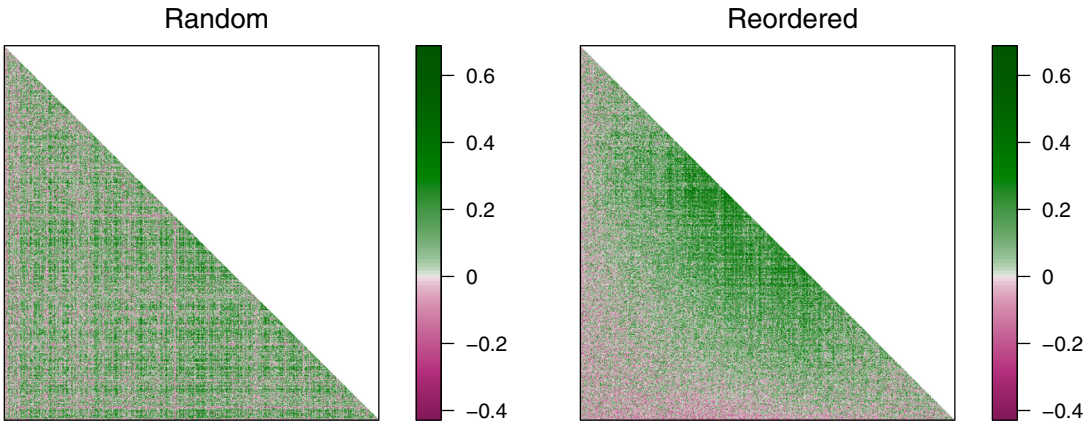

Figure 3. Empirical correlation matrix for the Chicago crime data. Left: random ordering of the variables. Right: matrix reordered through ARSA optimization

such ordering by defining the dissimilarity matrix $\mathbf{D}_e = \mathbf{1}_{N\times N} - \mathbf{R}_e$, where $\mathbf{1}_{N\times N}$ is the $N \times N$ matrix of ones, and appealing to the concept of anti-Robinson matrix (Hahsler *et al.*, 2008, sec. 2.1). In this type of matrix, the smallest dissimilarity (largest correlation) values appear close to the main diagonal and the largest dissimilarity (smallest correlation) values appear far from it. Hence, by defining a loss function that quantifies the divergence of a matrix from the anti-Robinson matrix (Hahsler *et al.*, 2008, sec. 2.2) reordering of the observations is solved by heuristic optimization employing the Anti-Robinson Simulated Annealing (ARSA); see Brusco *et al.* (2008). The `R` implementation of the algorithm is easily performed by using the package `seriation` (Hahsler *et al.*, 2008). The resulting ordering is quite satisfactorily and is plotted in Figure 3 (right) against a random ordering configuration (left).

Using this ordering the Gaussian AR-1 copula is selected 94% and 95% of the times, for the linear and the log-linear PNAR(1) model respectively. The estimated copula parameter is $\hat{\rho} = 0.689$ and $\hat{\rho} = 0.612$, for the linear and log-linear model respectively, with small SEs 0.064 and 0.062, correspondingly.

A further estimation step for the PNAR(1) models is performed by applying the two-step GEE estimation method discussed in Section S-7 in Supplement SM. The QMLE estimates are used as starting values of the two-step procedure. An AR-1 working correlation matrix $\mathbf{P}(\boldsymbol{\tau})$ is selected, with $\hat{\tau}_1$ as the estimator of the correlation parameter. To compare the relative efficiency of the GEE ($\tilde{\boldsymbol{\theta}}$) vs. QMLE ($\hat{\boldsymbol{\theta}}$), their bootstrap SEs have been calculated using 100 simulations by using the estimated copula. We compute the ratio of the SEs obtained, $q(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}) = \sum_{h=1}^{m} \mathrm{SE}(\hat{\beta}_h) / \sum_{h=1}^{m} \mathrm{SE}(\tilde{\beta}_h)$. The results are $q(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}) = 1.019$ and $q(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}) = 1.002$, for the linear and log-liner model respectively. We note a marginal gain in efficiency from the GEE estimation; this is probably due the a small value of the estimated correlation parameter τ, which is found to be around 0.008 and 0.005 on average, for linear and log-linear model respectively. Using different kind of estimator for the correlation parameter might yield significant efficiency improvement but a further study in this direction is needed.

ACKNOWLEDGEMENTS

This work was completed when M. Armillotta was with the Department of Mathematics & Statistics at the University of Cyprus. We greatly appreciate comments made by two reviewers on an earlier version of the manuscript. Both authors acknowledge the hospitality of the Department of Mathematics & Statistics at Lancaster University, where this work was initiated. This work has been co-financed by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation, under the project INFRASTRUCTURES/1216/0017 (IRIDA). In addition, K. Fokianos acknowledges travel support by CY Initiative of Excellence (grant 'Investissements d'Avenir' ANR-16-IDEX-0008), Project 'EcoDep' PSI-AAP2020-0000000013.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in GitHub at https://github.com/nick3703/Chicago-Data.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

REFERENCES

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, Francq C. 2016. Poisson QMLE of count time series models. *Journal of Time Series Analysis* **37**:291–314.

Aknouche A, Francq C. 2023. Two-stage weighted least squares estimator of the conditional mean of observation-driven time series models. *Journal of Econometrics* **237**:105174.

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

Alzaid A, Al-Osh M. 1990. An integer-valued pth-order autoregressive structure (INAR (p)) process. *Journal of Applied Probability* **27**:314–324.

Andreassen CM. 2013. Models and inference for correlated count data. Ph. D. thesis, Aaarhus University, Denmark.

Andrews DW. 1988. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* **4**:458–467.

Armillotta M, Luati A, Lupparelli M. 2022. Observation-driven models for discrete-valued time series. *Electronic Journal of Statistics* **16**:1393–1433.

Bracher J, Held L. 2022. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *International Journal of Forecasting* **38**:12211233.

Brusco MJ, Köhn H-F, Stahl S. 2008. Heuristic implementation of dynamic programming for matrix permutation problems in combinatorial data analysis. *Psychometrika* **73**:503–522.

Chen X, Chen Y, Xiao P. 2013. The impact of sampling and network topology on the estimation of social intercorrelations. *Journal of Marketing Research* **50**:95–110.

Chen X, Banks D, West M. 2019. Bayesian dynamic modeling and monitoring of network flows. *Network Science* **7**:292318.

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

Clark NJ, Dixon PM. 2021. A class of spatially correlated self-exciting statistical models. *Spatial Statistics* **43**:1–18.

Cliff A, Ord JK. 1975. Space-time modelling with an application to regional forecasting. *Transactions of the Institute of British Geographers* **64**:119–128.

Cox DR. 1981. Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics* **8**:93–115.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Systems* **1695**.

Cui Y, Zheng Q. 2017. Conditional maximum likelihood estimation for a class of observation-driven time series models for count data. *Statistics & Probability Letters* **123**:193–201.

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

Davis RA, Dunsmuir WTM, Streett SB. 2003. Observation-driven models for Poisson counts. *Biometrika* **90**:777–790.

Davis RA, Fokianos K, Holan SH, Joe H, Livsey J, Lund R, Pipiras V, Ravishanker N. 2021. Count time series: a methodological review. *Journal of the American Statistical Association* **116**:1533–1547.

Debaly ZM, Truquet L. 2019. Stationarity and moment properties of some multivariate count autoregressions. *arXiv preprint arXiv:1909.11392*.

Debaly ZM, Truquet L. 2021. A note on the stability of multivariate non-linear time series with an application to time series of counts. *Statistics & Probability Letters* **179**:1–7.

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

Douc R, Fokianos K, Moulines E. 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*. *Lecture Notes in Statistics*, Vol. 85 Springer-Verlag, New York.

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

DOI: 10.1111/jtsa.12728 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

Dunsmuir WT. 2016. *Generalized linear autoregressive moving average models*. In H*andbook of Discrete-Valued Time Series*, Vol. Chapter 3, Davis RA, Holan SH, Lund R, Ravishanker N (eds.). Chapman & Hall/CRC, London; 51–76.
Eyjolfsson H, Tjøstheim D. 2023. Multivariate self-exciting jump processes with applications to financial data. *Bernoulli* **29**:2167–2191.
Fang G, Xu G, Zhu X, Guan Y, et al. 2021. Group network Hawkes process. *arXiv preprint arXiv:2002.08521*.
Ferland R, Latour A, Oraichi D. 2006. Integer-valued GARCH process. *Journal of Time Series Analysis* **27**:923–942.
Fokianos K. 2021. Multivariate count time series modelling. *Econometrics and Statistics* To appear.
Fokianos K, Kedem B. 2004. Partial likelihood inference for time series following generalized linear models. *Journal of Time Series Analysis* **25**:173–197.
Fokianos K, Tjøstheim D. 2011. Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102**:563–578.
Fokianos K, Rahbek A, Tjøstheim D. 2009. Poisson autoregression. *Journal of the American Statistical Association* **104**:1430–1439.
Fokianos K, Støve B, Tjøstheim D, Doukhan P. 2020. Multivariate count autoregression. *Bernoulli* **26**:471–499.
Genest C, Nešlehová J. 2007. A primer on copulas for count data. *Astin Bulletin* **37**:475–515.
Gourieroux C, Monfort A, Trognon A. 1984. Pseudo maximum likelihood methods: Theory. *Econometrica* **52**:681–700.
Hahsler M, Hornik K, Buchta C. 2008. Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software* **25**:1–34.
Hall P, Heyde CC. 1980. *Martingale Limit Theory and its Application* Academic Press, Inc., New York.
Heinen A. 2003. Modelling time series count data: an autoregressive conditional Poisson model. Technical Report MPRA Paper 8113, University Library of Munich, Germany. http://mpra.ub.uni-muenchen.de/8113/.
Heinen A, Rengifo E. 2007. Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance* **14**:564–583.
Heyde CC. 1997. *Quasi-likelihood and its Application. A General Approach to Optimal Parameter Estimation*. Springer Series in Statistics Springer-Verlag, New York.
Inouye DI, Yang E, Allen GI, Ravikumar P. 2017. A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics* **9**:1–25.
Johnson SG. 2023. The NLopt nonlinear-optimization package. http://github.com/stevengj/nlopt.
Kedem B, Fokianos K. 2002. *Regression Models for Time Series Analysis* John Wiley & Sons, Hoboken, NJ.
Knight M, Nunes M, Nason G. 2016. Modelling, detrending and decorrelation of network time series. *arXiv:1603.03221*.
Knight M, Leeming K, Nason G, Nunes M. 2020. Generalized network autoregressive processes and the GNAR package. *Journal of Statistical Software* **96**:1–36.
Kolaczyk ED, Csárdi G. 2014. *Statistical Analysis of Network Data with R*, Vol. 65 Springer, New York.
Latour A. 1997. The multivariate GINAR (p) process. *Advances in Applied Probability* **29**:228–248.
Lee Y, Lee S, Tjøstheim D. 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.
Martin RL, Oeppen J. 1975. The identification of regional forecasting models using space: time correlation functions. *Transactions of the Institute of British Geographers* **66**:95–118.
McCullagh P, Nelder JA. 1989. *Generalized Linear Models*, 2nd ed. Chapman & Hall, London.
Meyn SP, Tweedie RL. 1993. *Markov Chains and Stochastic Stability* Springer, London.
Neumann M. 2011. Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17**:1268–1284.
Nowicki K, Snijders TAB. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**:1077–1087.
Pan W. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics* **57**:120–125.
Pedeli X, Karlis D. 2011. A bivariate INAR(1) process with application. *Statistical Modelling* **11**:325–349.
Pedeli X, Karlis D. 2013a. On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis* **34**:206–220.
Pedeli X, Karlis D. 2013b. Some properties of multivariate INAR(1) processes. *Computational Statistics & Data Analysis* **67**:213–225.
Pfeifer PE, Deutrch SJ. 1980. A three-stage iterative procedure for space-time modeling. *Technometrics* **22**:35–47.
Pötscher BM, Prucha IR. 1997. *Dynamic Nonlinear Econometric Models*. Asymptotic theory Springer-Verlag, Berlin.
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.
Veraart AE. 2019. Modeling, simulation and inference for multivariate time series of counts using trawl processes. *Journal of Multivariate Analysis* **169**:110–129.
Wang YJ, Wong GY. 1987. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association* **82**:8–19.

Wang C, Liu H, Yao J-F, Davis RA, Li WK. 2014. Self-excited threshold Poisson autoregression. *Journal of the American Statistical Association* **109**:777–787.
Wasserman S, Faust K, et al. 1994. *Social Network Analysis: Methods and Applications*, Vol. 8 Cambridge University Press, Cambridge.
Wedderburn RW. 1974. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**:439–447.
Weiß CH. 2018. *An Introduction to Discrete-valued Time Series* John Wiley & Sons, Hoboken, NJ.
West M. 2020. Bayesian forecasting of multivariate time series: scalability, structure uncertainty and decisions. *Annals of the Institute of Statistical Mathematics* **72**:1–31.
West M, Harrison P. 1997. *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer, New York.
Woodard DW, Matteson DS, Henderson SG. 2011. Stationarity of count-valued and nonlinear time series models. *Electronic Journal of Statistics* **5**:800–828.
Zeger SL. 1988. A regression model for time series of counts. *Biometrika* **75**:621–629.
Zeger SL, Liang K-Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**:121–130.
Zhang C, Chen N, Li Z. 2017. State space modeling of autocorrelated multivariate Poisson counts. *IISE Transactions* **49**:518–531.
Zhao Y, Levina E, Zhu J, et al. 2012. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* **40**(4):2266–2292.
Zhou J, Li D, Pan R, Wang H. 2020. Network GARCH model. *Statistica Sinica* **30**:1–18.
Zhu X, Pan R. 2020. Grouped network vector autoregression. *Statistica Sinica* **30**:1437–1462.
Zhu X, Pan R, Li G, Liu Y, Wang H. 2017. Network vector autoregression. *The Annals of Statistics* **45**:1096–1123.
Zhu X, Wang W, Wang H, Härdle WK. 2019. Network quantile autoregression. *Journal of Econometrics* **212**:345–358.

APPENDIX A: APPENDIX

Recall that C is a generic constant and C_r is a constant depending on $r \in \mathbb{N}$. See also the notation paragraph in the introductory Section 1.

A.1. *Proof of Theorem 1*

Recall from Zhu *et al.* (2017, def. 1) that $\mathcal{W} = \left\{\boldsymbol{\omega} \in \mathbb{R}^\infty : \boldsymbol{\omega}_\infty = \sum |\omega_i| < \infty\right\}$, where $\boldsymbol{\omega} = (\omega_i \in \mathbb{R} : 1 \leq i < \infty)^T \in \mathbb{R}^\infty$. For each $\boldsymbol{\omega} \in \mathcal{W}$, let $\boldsymbol{\omega}_N = (\omega_1, \dots, \omega_N)^T \in \mathbb{R}^N$ be the its truncated N-dimensional version. By considering the VAR(1) representation for the PNAR(1) model (2), defined in Section S-1 in Supplement SM, the process can be rewritten by backward substitution, $\mathbf{Y}_t = (\mathbf{I}_N - \mathbf{G})^{-1}\boldsymbol{\beta}_0 + \sum_{j=0}^{\infty}\mathbf{G}^j\boldsymbol{\xi}_{t-j}$. For sake of clarity we show the result for the PNAR(1) model. However, the general p-lags parallel result extends straightforwardly, by considering the companion VAR(1) representation form (Section S-1 in Supplement SM) of the linear PNAR(p) model. By Proposition 2, it holds that $\mathrm{E}(Y_{i,t}) \leq \mu = \beta_0/(1-\beta_1-\beta_2)$ for all $1 \leq i < \infty$ and, since $\boldsymbol{\xi}_t = \mathbf{Y}_t - \boldsymbol{\lambda}_t$, $\mathrm{E}\left|\xi_{i,t}\right| \leq 2\mathrm{E}(Y_{i,t}) \leq 2\mu = c < \infty$. Similar uniform bounds are obtained for moments of order $r > 1$. For any $\boldsymbol{\omega} \in \mathcal{W}$, $\mathrm{E}\left|\boldsymbol{\beta}_0 + \boldsymbol{\xi}_t\right|_v \leq (\beta_0 + c)\mathbf{1}_N = C\mathbf{1}_N < \infty$, $\mathbf{G}^j\mathbf{1}_N = (\beta_1+\beta_2)^j\mathbf{1}_N$ and $\mathrm{E}\left|\boldsymbol{\omega}_N^T\sum_{j=0}^{\infty}\mathbf{G}^j(\boldsymbol{\beta}_0 + \boldsymbol{\xi}_{t-j})\right| \leq C\omega_\infty\sum_{j=0}^{\infty}(\beta_1+\beta_2)^j = C_2$. Then, by Monotone Convergence Theorem (MCT), $\lim_{N\to\infty}\boldsymbol{\omega}_N^T\mathbf{Y}_t$ exists and is finite with probability 1, moreover $Y_t^\omega = \lim_{N\to\infty}\boldsymbol{\omega}_N^T\mathbf{Y}_t$ is strictly stationary and therefore $\left\{\mathbf{Y}_t\right\}$ is strictly stationary, following Zhu *et al.* (2017, def. 1). To verify the uniqueness of the solution, take another stationary solution $\tilde{\mathbf{Y}}_t$ to the PNAR model with finite moments of any order. Then, $\mathrm{E}(\tilde{\mathbf{Y}}_t) \leq C_1\mathbf{1}_N$, where C_1 is a constant and $\mathrm{E}\left|\boldsymbol{\omega}_N^T\mathbf{Y}_t - \boldsymbol{\omega}_N^T\tilde{\mathbf{Y}}_t\right| = \left|\sum_{j=m}^{\infty}\boldsymbol{\omega}_N^T\sum_{j=0}^{\infty}\mathbf{G}^j(\boldsymbol{\beta}_0 + \boldsymbol{\xi}_{t-j}) - \boldsymbol{\omega}_N^T\mathbf{G}^m\tilde{\mathbf{Y}}_{t-m}\right| \leq \omega_\infty\sum_{j=m}^{\infty}[C_2(\beta_1+\beta_2)^j + C_1(\beta_1+\beta_2)^m]$, for any N and weight $\boldsymbol{\omega}$. Since m is arbitrary, $Y_t^\omega = \tilde{Y}_t^\omega$ with probability one. □

A.2. *Proof of Lemma 1*

We split the proof accordingly to each single result given in Lemma 1.

Proof of (1). Define $\mathbf{W}_t = (\mathbf{Y}_t, \mathbf{Y}_{t-1})^T$, $\hat{\mathbf{W}}_{t-J}^t = \left(\hat{\mathbf{Y}}_{t-J}^t, \hat{\mathbf{Y}}_{t-J}^{t-1}\right)^T := f(\boldsymbol{\xi}_t, \dots, \boldsymbol{\xi}_{t-J})$, $\hat{Y}_{i,t}$, $\hat{\lambda}_{i,t}$ the ith elements of $\hat{\mathbf{Y}}_{t-J}^t$ and $\hat{\boldsymbol{\lambda}}_{t-J}^t$. Consider the following triangular array $\left\{g_{Nt}(\mathbf{W}_t) : 1 \leq t \leq T_N, N \geq 1\right\}$, where $T_N \to \infty$ as

DOI: 10.1111/jtsa.12728

$N \to \infty$. For any $\boldsymbol{\eta} \in \mathbb{R}^m$, $g_{Nt}(\mathbf{W}_t) = N^{-1}\boldsymbol{\eta}^T \frac{\partial \lambda_t^T}{\partial \boldsymbol{\theta}} \mathbf{C}_t \frac{\partial \lambda_t}{\partial \boldsymbol{\theta}^T} \boldsymbol{\eta} = \sum_{r=1}^m \sum_{l=1}^m \eta_r \eta_l h_{rl,t}$ where $N^{-1}\mathbf{H}_{Nt} = (h_{rl,t})_{1\le r,l\le m}$. We take the most complicated element, $h_{22,t}$, the result is analogously proven for the other elements. Define $l_{1,i,t} = \left|(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 Y_{i,t}(\hat{\lambda}_{i,t} + \lambda_{i,t})\right|$, $l_{2,i,t} = \left|(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 \lambda_{i,t}^2\right|$ and $l_{3,i,t} = \left|\hat{Y}_{i,t}\lambda_{i,t}^2 (Y_{i,t-1} + \hat{Y}_{i,t-1}) \sum_{j=1}^N w_{ij}(Y_{j,t-1} + \hat{Y}_{j,t-1})\right|$. Additionally, the equality $\left|\hat{\lambda}_{i,t} - \lambda_{i,t}\right| = \left|Y_{i,t} - \hat{Y}_{i,t}\right|$ is a consequence of the constructions in Lemma S2 in Supplement SM. Then

$$
\begin{aligned}
\left|h_{22,t} - h_{22,t-J}^t\right| &= \left|\frac{1}{N}\sum_{i=1}^N \frac{(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 Y_{i,t}}{\lambda_{i,t}^2} - \frac{1}{N}\sum_{i=1}^N \frac{(\mathbf{w}_i^T \hat{\mathbf{Y}}_{t-J}^{t-1})^2 \hat{Y}_{i,t}}{\hat{\lambda}_{i,t}^2}\right| \\
&\le \frac{\beta_0^{-4}}{N}\sum_{i=1}^N \left|(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 Y_{i,t}\hat{\lambda}_{i,t}^2 - (\mathbf{w}_i^T \hat{\mathbf{Y}}_{t-J}^{t-1})^2 \hat{Y}_{i,t}\lambda_{i,t}^2\right| \\
&\le \frac{\beta_0^{-4}}{N}\sum_{i=1}^N \left|(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 Y_{i,t}(\hat{\lambda}_{i,t}^2 - \lambda_{i,t}^2) + \left[(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 Y_{i,t} - (\mathbf{w}_i^T \hat{\mathbf{Y}}_{t-J}^{t-1})^2 \hat{Y}_{i,t}\right] \lambda_{i,t}^2\right| \\
&\le \frac{\beta_0^{-4}}{N}\left|\sum_{i=1}^N (\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 Y_{i,t}(\hat{\lambda}_{i,t} + \lambda_{i,t})(\hat{\lambda}_{i,t} - \lambda_{i,t})\right| + \frac{\beta_0^{-4}}{N}\left|\sum_{i=1}^N (\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 \lambda_{i,t}^2 (Y_{i,t} - \hat{Y}_{i,t})\right| \\
&\quad + \frac{\beta_0^{-4}}{N}\left|\sum_{i=1}^N \hat{Y}_{i,t}\lambda_{i,t}^2 \left[\left(\mathbf{w}_i^T \mathbf{Y}_{t-1}\right)^2 - \left(\mathbf{w}_i^T \hat{\mathbf{Y}}_{t-J}^{t-1}\right)^2\right]\right| \\
&\le \frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{1it}\left|\hat{\lambda}_{i,t} - \lambda_{i,t}\right| + \frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{2it}\left|Y_{i,t} - \hat{Y}_{i,t}\right| \\
&\quad + \frac{\beta_0^{-4}}{N}\sum_{i=1}^N \hat{Y}_{i,t}\lambda_{i,t}^2 \left|\left(\mathbf{w}_i^T \mathbf{Y}_{t-1}\right) + \left(\mathbf{w}_i^T \hat{\mathbf{Y}}_{t-J}^{t-1}\right)\right| \left|\left(\mathbf{w}_i^T \mathbf{Y}_{t-1}\right) - \left(\mathbf{w}_i^T \hat{\mathbf{Y}}_{t-J}^{t-1}\right)\right| \\
&\le \frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{1it}\left|\hat{\lambda}_{i,t} - \lambda_{i,t}\right| + \frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{2it}\left|Y_{i,t} - \hat{Y}_{i,t}\right| \\
&\quad + \frac{\beta_0^{-4}}{N}\sum_{i=1}^N \hat{Y}_{i,t}\lambda_{i,t}^2 \left|\sum_{j=1}^N w_{ij}\left(Y_{j,t-1} + \hat{Y}_{j,t-1}\right)\right| \left|\sum_{j=1}^N w_{ij}\left(Y_{j,t-1} - \hat{Y}_{j,t-1}\right)\right| \\
&\le \frac{\beta_0^{-4}}{N}\sum_{i=1}^N \left(l_{1,i,t} + l_{2,i,t}\right)\left|Y_{i,t} - \hat{Y}_{i,t}\right| + \frac{\beta_0^{-4}}{N}\sum_{i=1}^N l_{3,i,t}\left|\sum_{j=1}^N w_{ij}\left(Y_{j,t-1} - \hat{Y}_{j,t-1}\right)\right| .
\end{aligned}
$$

Set $1/a + 1/b = 1/2$ and $1/q + 1/p + 1/n = 1/a$. By Cauchy–Schwartz inequality, as $w_{ij} > 0$ for $j = 1, \dots, N$ and $\sum_{j=1}^N w_{ij} = 1$ we have that $(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2 = \left(\sum_{j=1}^N w_{ij} Y_{j,t-1}\right)^2 \le \sum_{j=1}^N w_{ij} Y_{j,t-1}^2$. As a consequence, $\max_{1\le i\le N} \left\|(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2\right\|_q \le \max_{1\le i\le N}\left(\sum_{j=1}^N w_{ij}\left\|Y_{j,t-1}^2\right\|_q\right) \le \sup_{i\ge 1}\left\|Y_{i,t}^2\right\|_q \le C_{2q}^{1/q} < \infty$, by Proposition 2. Moreover, $\sup_{i\ge 1}\left\|\lambda_{i,t}^2\right\|_n \le \sup_{i\ge 1}\left\|Y_{i,t}^2\right\|_n \le C_n$, by the conditional Jensen's inequality. Similarly, $\sup_{i\ge 1}\left\|\hat{\lambda}_{i,t}^2\right\|_n \le \sup_{i\ge 1}\left\|\hat{Y}_{i,t}^2\right\|_n$. An application of Lemma S-2 in Supplement SM provides $\sup_{i\ge 1}\left\|Y_{i,t} - \hat{Y}_{i,t}\right\|_b \le d^J \sum_{j=0}^{t-J-1} d^j \sup_{i\ge 1}\left\|\xi_{i,t}\right\|_b \le d^J 2C_b^{1/b}/(1-d)$. By an analogous recursion argument, it holds that $\sup_{i\ge 1}\left\|\hat{Y}_{i,t}^2\right\|_n \le 2\beta_0 \sum_{j=0}^\infty d^j + \sum_{j=0}^\infty d^j \sup_{i\ge 1}\left\|\xi_{i,t}\right\|_n \le (2\beta_0 + 2C_n^{1/n})/(1-d) := \Delta < \infty$. It is immediate to see that, by Holder's inequality $\sup_{i\ge 1}\left\|l_{1,i,t}\right\|_a \le \sup_{i\ge 1}\left\|(\mathbf{w}_i^T \mathbf{Y}_{t-1})^2\right\|_q \left\|Y_{i,t}\right\|_p \left(\left\|\hat{\lambda}_{i,t}\right\|_n + \left\|\hat{\lambda}_{i,t}\right\|_n\right) < l_1 < \infty$. In the same way we can

conclude that $\sup_{i\geq 1}\left\|l_{2,i,t}\right\|_q < l_2 < \infty$ and $\sup_{i\geq 1}\left\|l_{3,i,t}\right\|_q < l_3 < \infty$. Then, by Minkowski inequality

$$\begin{aligned}\left\|h_{22,t} - h^t_{22,t-J}\right\|_2 &\leq \frac{\beta_0^{-4}}{N}\sum_{i=1}^{N}\left\|l_{1,i,t}+l_{2,i,t}\right\|_a\left\|Y_{i,t}-\hat{Y}_{i,t}\right\|_b + \frac{\beta_0^{-4}}{N}\sum_{i=1}^{N}\left\|l_{3,i,t}\right\|_a\sum_{j=1}^{N}w_{ij}\left\|Y_{j,t-1}-\hat{Y}_{j,t-1}\right\|_b \\ &\leq \beta_0^{-4}\max_{1\leq i\leq N}\left(\left\|l_{1,i,t}\right\|_a+\left\|l_{2,i,t}\right\|_a\right)\left\|Y_{i,t}-\hat{Y}_{i,t}\right\|_b + \beta_0^{-4}\max_{1\leq i\leq N}\left\|l_{3,i,t}\right\|_a\left\|Y_{i,t-1}-\hat{Y}_{i,t-1}\right\|_b \\ &\leq \beta_0^{-4}\left(l_1+l_2+l_3\right)2C_b^{1/b}d^{J-1}\sum_{j=0}^{t-J-1}d^j \leq \frac{\beta_0^{-4}\left(l_1+l_2+l_3\right)2C_b^{1/b}}{1-d}d^{J-1} := c_{22}\nu_J\,,\end{aligned}$$

with $\nu_J = d^{J-1}$. By the definition in B1, set $\mathcal{F}^N_{t-J,t+J} = \sigma\left(\xi_{i,t} : 1\leq i\leq N, t-J\leq t\leq t+J\right)$. 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 $\left\{\overline{X}_{Nt} = g_{Nt}(\mathbf{W}_t) - \mathrm{E}\left[g_{Nt}(\mathbf{W}_t)\right]\right\}$; see Andrews (1988). Moreover, by a similar argument above, it is easy to see that $\mathrm{E}\left|\overline{X}_{Nt}\right|^2 < \infty$, by the finiteness of all the moments of the process $\mathbf{Y}_t$. Then, using B1 and the argument in Andrews (1988, p. 464), we have that $\left\{\overline{X}_{Nt}\right\}$ 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 $(NT_N)^{-1}\boldsymbol{\eta}^T\mathbf{H}_{NT_N}\boldsymbol{\eta}\xrightarrow{p}\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}$ as $\left\{N,T_N\right\}\to\infty$. We only need to show the existence of the matrix $\mathbf{H}$ according to (20). Consider the single elements of the matrix $\mathbf{H}_N$:

$$H_{11} = \mathrm{E}\left(\sum_{i=1}^{N}\frac{1}{\lambda_{i,t}}\right),\quad H_{12} = \mathrm{E}\left(\sum_{i=1}^{N}\frac{\mathbf{w}_i^T\mathbf{Y}_{t-1}}{\lambda_{i,t}}\right),\quad H_{13} = \mathrm{E}\left(\sum_{i=1}^{N}\frac{Y_{i,t-1}}{\lambda_{i,t}}\right),$$

$$H_{22} = \mathrm{E}\left[\sum_{i=1}^{N}\frac{\left(\mathbf{w}_i^T\mathbf{Y}_{t-1}\right)^2}{\lambda_{i,t}}\right],\quad H_{23} = \mathrm{E}\left(\sum_{i=1}^{N}\frac{\mathbf{w}_i^T\mathbf{Y}_{t-1}Y_{i,t-1}}{\lambda_{i,t}}\right),$$

$$H_{33} = \mathrm{E}\left[\sum_{i=1}^{N}\frac{\left(Y_{i,t-1}\right)^2}{\lambda_{i,t}}\right].$$

Note that the linear model (2) can be rewritten has $\mathbf{Y}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \mathbf{G}^j \boldsymbol{\xi}_{t-j} = \boldsymbol{\mu} + \tilde{\mathbf{Y}}_t$ where $\boldsymbol{\mu} = (\mathbf{I}_N - \mathbf{G})^{-1}\beta_0 \mathbf{1} = \beta_0(1-\beta_1-\beta_2)^{-1}\mathbf{1}$ and $\boldsymbol{\xi}_t$ is MDS. As $N \to \infty$,

$$\frac{1}{N} H_{11} = \mathrm{E}\left(\frac{1}{N} \mathbf{1}^T \mathbf{D}_t^{-1} \mathbf{1}\right) = \frac{1}{N} \mathrm{tr}(\boldsymbol{\Lambda}) \to d_1, \tag{A-1}$$

by assumption B3. The second term

$$\frac{1}{N} H_{12} = \mathrm{E}\left(\frac{1}{N} \mathbf{1}^T \mathbf{D}_t^{-1} \mathbf{W} \mathbf{Y}_{t-1}\right) = \frac{1}{N} H_{12a} + \frac{1}{N} H_{12b},$$

where $H_{12a}/N = N^{-1}\mathrm{E}\left(\mathbf{1}^T \mathbf{D}_t^{-1} \mathbf{W} \boldsymbol{\mu}\right) = N^{-1} \mathbf{1}^T \boldsymbol{\Lambda} \mathbf{W} (\mathbf{I}_N - \mathbf{G})^{-1} \beta_0 \mathbf{1} = \beta_0 N^{-1} \mathbf{1}^T \boldsymbol{\Lambda} \mathbf{W} (1-\beta_1-\beta_2)^{-1} \mathbf{1} = \mu \mathbf{1}^T \boldsymbol{\Lambda} \mathbf{W} \mathbf{1}/N = \mu \mathbf{1}^T \boldsymbol{\Lambda} \mathbf{1}/N \to \mu d_1$, as $N \to \infty$. Define $e_{i,t} = \left| \boldsymbol{\xi}_{t-1-i}^T \right|_v (\mathbf{G}^T)^i \mathbf{W}^T \mathbf{1}$. Then,

$$\begin{aligned}
\left|\frac{H_{12b}}{N}\right| &\le \frac{1}{N}\left[\mathrm{E}\left(\mathbf{1}^T \mathbf{D}_t^{-1} \mathbf{W} \tilde{\mathbf{Y}}_{t-1}\right)^2\right]^{1/2} \le \frac{\beta_0^{-1}}{N}\left[\mathrm{E}\left(\mathbf{1}^T \mathbf{W} \left|\tilde{\mathbf{Y}}_{t-1}\right|_v\right)^2\right]^{1/2} \\
&\le \frac{\beta_0^{-1}}{N} \sum_{i,j=0}^{\infty} \mathrm{E}^{1/2}\left(\mathbf{1}^T \mathbf{W} \mathbf{G}^j \left|\boldsymbol{\xi}_{t-1-j}\right|_v \left|\boldsymbol{\xi}_{t-1-i}^T\right|_v (\mathbf{G}^T)^i \mathbf{W}^T \mathbf{1}\right) \\
&\le \frac{\beta_0^{-1}}{N} \sum_{i,j=0}^{\infty} \mathrm{E}^{1/4}(e_{j,t}^2)\mathrm{E}^{1/4}(e_{i,t}^2) = \frac{\beta_0^{-1}}{N}\left[\sum_{j=0}^{\infty} \mathrm{E}^{1/4}(e_{j,t}^2)\right]^2 \\
&\le \beta_0^{-1}\left[\sum_{j=0}^{\infty} \frac{1}{\sqrt{N}} \mathrm{E}^{1/4}\left(\mathbf{1}^T \mathbf{W} \mathbf{G}^j \left|\boldsymbol{\xi}_{t-1-j} \boldsymbol{\xi}_{t-1-j}^T\right|_v (\mathbf{G}^T)^j \mathbf{W}^T \mathbf{1}\right)\right]^2 \\
&\le \beta_0^{-1}\left[\sum_{j=0}^{\infty} \frac{1}{\sqrt{N}} \left(\mathbf{1}^T \mathbf{W} \mathbf{G}^j \boldsymbol{\Sigma}_{\boldsymbol{\xi}} (\mathbf{G}^T)^j \mathbf{W}^T \mathbf{1}\right)^{1/4}\right]^2,
\end{aligned} \tag{A-2}$$

converges to 0, as $N \to \infty$, where the first inequality holds by Minkowski and Jensen's inequalities, the second inequality is a consequence of $\mathbf{D}_t^{-1} \preceq \beta_0^{-1}\mathbf{I}_N$ and the fourth is deduced by Cauchy inequality. The convergence follows by applying Lemma S1 in Supplement SM. Then, $H_{12}/N \to \mu d_1$ as $N \to \infty$. For the same reason $H_{13}/N \to \mu d_1$. We move to the following term.

$$\frac{H_{22}}{N} = \mathrm{E}\left(\frac{1}{N} \mathbf{Y}_{t-1}^T \mathbf{W}^T \mathbf{D}_t^{-1} \mathbf{W} \mathbf{Y}_{t-1}\right) = \frac{H_{22a}}{N} + \frac{H_{22b}}{N} + \frac{H_{22c}}{N} + \frac{H_{22d}}{N},$$

where, as $N \to \infty$, $H_{22a}/N = \mathrm{E}\left(N^{-1} \boldsymbol{\mu}^T \mathbf{W}^T \mathbf{D}_t^{-1} \mathbf{W} \boldsymbol{\mu}\right) = \mu^2 \mathbf{1}^T \mathbf{W}^T \boldsymbol{\Lambda} \mathbf{W} \mathbf{1}/N = \mu^2 \mathrm{tr}(\boldsymbol{\Lambda})/N \to \mu^2 d_1$ and $H_{22b}/N = H_{22c}/N = \mu H_{12b}/N \to 0$. Finally,

$$\frac{H_{22d}}{N} = \frac{1}{N}\mathrm{E}\left(\tilde{\mathbf{Y}}_{t-1}^T \mathbf{W}^T \mathbf{D}_t^{-1} \mathbf{W} \tilde{\mathbf{Y}}_{t-1}\right) = \frac{1}{N}\mathrm{tr}\ \ \mathrm{E}\left[\mathbf{D}_t^{-1/2} \mathbf{W} \left(\mathbf{Y}_{t-1} - \boldsymbol{\mu}\right)\left(\mathbf{Y}_{t-1} - \boldsymbol{\mu}\right)^T \mathbf{W}^T \mathbf{D}_t^{-1/2}\right] \to d_4,$$

as $N \to \infty$, using B3. So $H_{22}/N \to \mu^2 d_1 + d_4$ as $N \to \infty$. For the same reason $H_{23}/N \to \mu^2 d_1 + d_3$ and $H_{33}/N \to \mu^2 d_1 + d_2$. Finally, note that $\mathbf{H}$ is positive definite, and nonsingular, as $\mathbf{H}_N/N$ is positive definite. □

Proof of (2). For all non-null $\boldsymbol{\eta} \in \mathbb{R}^m$, the triangular array $\left\{\boldsymbol{\eta}^T \mathbf{s}_{Nt}/N : 1 \le t \le T_N, N \ge 1\right\}$ is a martingale difference array. Moreover, $\mathrm{E}(\boldsymbol{\eta}^T \mathbf{s}_{Nt}/N)^2 < \infty$, by Cauchy inequality and the boundedness of all the moments of

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728

$\{\mathbf{Y}_t\}$. Then, $\boldsymbol{\eta}^T\mathbf{s}_{Nt}/N$ is trivially a uniformly integrable L^1-mixingale. An application of Andrews (1988, thm. 2) provides the result. □

Proof of (3). From $\theta \in \mathcal{O}(\theta_0)$, we have $\beta_{0,*} \leq \beta_0 \leq \beta_0^*$, where $\beta_{0,*}, \beta_0^*$ are suitable positive constants. Consider the third derivative

$$\frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_l\partial\theta_k} = 2\frac{Y_{i,t}}{\lambda_{i,t}^3(\boldsymbol{\theta})}\left(\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\theta_k}\right) \leq 2\frac{\beta_{0,*}^{-1}Y_{i,t}}{\lambda_{i,t}^2(\boldsymbol{\theta})}\left(\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\theta_{j^*}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\theta_{l^*}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\theta_{k^*}}\right) := m_{i,t}.$$

Take the maximum of the third derivatives among $\{i,l,k\}$ to be, for example, at $\theta_{j^*} = \theta_{l^*} = \theta_{k^*} = \beta_1$, the proof is analogous for the other derivatives,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^3 l_{i,t}(\boldsymbol{\theta})}{\partial\beta_1^3} = \frac{1}{N}\sum_{i=1}^{N}2\frac{Y_{i,t}}{\lambda_{i,t}^3(\boldsymbol{\theta})}\left(\mathbf{w}_i^T\mathbf{Y}_{t-1}\right)^3 \leq \frac{1}{N}\sum_{i=1}^{N}2\frac{\beta_{0,*}^{-1}Y_{i,t}}{\lambda_{i,t}^2(\boldsymbol{\theta})}\left(\mathbf{w}_i^T\mathbf{Y}_{t-1}\right)^3 := \frac{1}{N}\sum_{i=1}^{N}m_{i,t}.$$

Now, define $M_{NT_N} := (NT_N)^{-1}\sum_{t=1}^{T_N}\sum_{i=1}^{N}m_{i,t}$ and $N^{-1}\sum_{i=1}^{N}\mathrm{E}(m_{i,t}) < \infty$ since all the moment of $\mathbf{Y}_t$ exist. It is easy to see that $M_{NT_N} \xrightarrow{p} M$ as $\{N,T_N\} \to \infty$, similarly to the steps of A.2.1 above, with $M = \lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\mathrm{E}(m_{i,t})$. Then point (3) of Lemma 1 follows by the last limit of B3. We omit the details. □

A.3. Proof of Lemma 2

Analogously to A.2, we address separately each point of Lemma 2.

Proof of (1). Let $\tilde{g}_{Nt}(\mathbf{W}_t) = N^{-1}\boldsymbol{\eta}^T\frac{\partial\boldsymbol{\lambda}_t^T}{\partial\boldsymbol{\theta}}\mathbf{D}_t^{-1}\boldsymbol{\Sigma}_t\mathbf{D}_t^{-1}\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}^T}\boldsymbol{\eta} = \sum_{r=1}^{m}\sum_{l=1}^{m}\eta_r\eta_l b_{rl,t}$ where $N^{-1}\mathbf{B}_{Nt} = (b_{rl,t})_{1\leq r,l\leq m}$ and $\boldsymbol{\Sigma}_t = \mathrm{E}(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^T|\mathcal{F}_{t-1}^N)$, with $\boldsymbol{\xi}_t = \mathbf{Y}_t - \boldsymbol{\lambda}_t = \hat{\mathbf{Y}}_{t-J}^t - \hat{\boldsymbol{\lambda}}_{t-J}^t$, since $\mathrm{E}\left(\hat{\mathbf{Y}}_{t-J}^t|\mathcal{F}_{t-1}^N\right) = \hat{\boldsymbol{\lambda}}_{t-J}^t$. We consider again the most complicated element, that is $b_{22,t}$. For $1 \leq i,j \leq N$, define $\sigma_{ijt} = \mathrm{E}\left(\xi_{i,t}\xi_{j,t}|\mathcal{F}_{t-1}^N\right)$ and $\rho_{ijt} = \mathrm{E}\left(\xi_{i,t}\xi_{j,t}|\mathcal{F}_{t-1}^N\right) / \left(\sqrt{\lambda_{i,t}}\sqrt{\lambda_{j,t}}\right)$, which are the elementwise conditional covariances and correlations respectively. Then

$$\begin{aligned}
\left|b_{22,t} - b_{22,t-J}^t\right| &= \left|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\left(\mathbf{w}_i^T\mathbf{Y}_{t-1}\right)\left(\mathbf{w}_j^T\mathbf{Y}_{t-1}\right)\sigma_{ijt}}{\lambda_{i,t}\lambda_{j,t}} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\left(\mathbf{w}_i^T\hat{\mathbf{Y}}_{t-J}^{t-1}\right)\left(\mathbf{w}_j^T\hat{\mathbf{Y}}_{t-J}^{t-1}\right)\sigma_{ijt}}{\hat{\lambda}_{i,t}\hat{\lambda}_{j,t}}\right| \\
&\leq \beta_0^{-3}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\left|\sigma_{ijt}\right|}{\lambda_{i,t}^{1/2}\lambda_{j,t}^{1/2}}\left|\left(\mathbf{w}_i^T\mathbf{Y}_{t-1}\right)\left(\mathbf{w}_j^T\mathbf{Y}_{t-1}\right)\hat{\lambda}_{i,t}\hat{\lambda}_{j,t} - \left(\mathbf{w}_i^T\hat{\mathbf{Y}}_{t-J}^{t-1}\right)\left(\mathbf{w}_j^T\hat{\mathbf{Y}}_{t-J}^{t-1}\right)\lambda_{i,t}\lambda_{j,t}\right| \\
&\leq \beta_0^{-3}\frac{1}{N}\sum_{i,j=1}^{N}\left|\rho_{ijt}\right|\left(r_{1,i,j,t}\left|\lambda_{i,t} - \hat{\lambda}_{i,t}\right| + r_{2,i,j,t}\left|\sum_{j=1}^{N}w_{ij}(Y_{j,t-1} - \hat{Y}_{j,t-1})\right|\right).
\end{aligned}$$

The second inequality is obtained employing the arguments used for the element $h_{22,t}$ of the Hessian as in A.2. Moreover, $r_{1,i,j,t} = (\mathbf{w}_i^T\mathbf{Y}_{t-1})(\mathbf{w}_j^T\mathbf{Y}_{t-1})(\hat{\lambda}_{j,t} + \lambda_{j,t})$ and $r_{2,i,j,t} = \lambda_{i,t}\lambda_{j,t}(\mathbf{w}_i^T\mathbf{Y}_{t-1} + \mathbf{w}_i^T\mathbf{Y}_{t-J}^{t-1})$. Set $1/q + 1/h = 1/b$. Note that $\sup_{i,j\geq 1}\left\|r_{1,i,j,t}\right\|_q < r_1 < \infty$, $\sup_{i,j\geq 1}\left\|r_{2,i,j,t}\right\|_q < r_2 < \infty$ by the same argument of $\sup_{i\geq 1}\left\|l_{1,i,t}\right\|_a < l_1$ above. When $i = j$, $\sigma_{iit} = \lambda_{i,t}$, consequently, $N^{-1}\sum_{i,j=1}^{N}\left\|\rho_{ijt}\right\|_a = N^{-1}\sum_{i=1}^{N}\|1\|_a = 1$. Instead, when $i \neq j$,

$$\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|\rho_{ijt}\right| = \max_{1\leq i\leq N}\sum_{j=1}^{i-1}\left|\rho_{ijt}\right| + \max_{1\leq i\leq N}\sum_{j=i+1}^{N}\left|\rho_{ijt}\right| \leq \max_{1\leq i\leq N}\sum_{j=1}^{i-1}\varphi_{i-j} + \max_{1\leq i\leq N}\sum_{j=i+1}^{N}\varphi_{j-i} \leq 2\sum_{h=1}^{N-1}\varphi_h$$

which is bounded by 2Φ and the first inequality is a consequence of B4. Then, $\forall i,j = 1,\dots,N$, we have $N^{-1}\sum_{i,j=1}^{N}\left\|\rho_{ijt}\right\|_a \leq \lambda$, where $\lambda = \max\{1, 2\Phi\}$. This entails that

$$\begin{aligned}\left\|b_{22,t} - b^t_{22,t-J}\right\|_2 &\leq \beta_0^{-3}\frac{1}{N}\sum_{i,j=1}^{N}\left\|\rho_{ijt}\right\|_a\left\|r_{1,i,j,t}\left|\lambda_{i,t}-\hat{\lambda}_{i,t}\right| + r_{2,i,j,t}\left|\sum_{j=1}^{N} w_{ij}(Y_{j,t-1}-\hat{Y}_{j,t-1})\right|\right\|_b \\ &\leq \beta_0^{-3}\lambda\max_{1\leq i,j\leq N}\left\|r_{1,i,j,t}\right\|_q\left\|Y_{i,t}-\hat{Y}_{i,t}\right\|_h + \beta_0^{-4}\lambda\max_{1\leq i,j\leq N}\left\|r_{2,i,j,t}\right\|_q\left\|Y_{i,t-1}-\hat{Y}_{i,t-1}\right\|_h \\ &\leq \frac{\beta_0^{-3}\lambda\left(r_1+r_2\right)2C_h^{1/h}}{1-d}d^{J-1} := r_{22}\nu_J\,.\end{aligned}$$

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, with $\mathrm{E}\tilde{X}^2_{Nt} < \infty$, and theorem 2 in Andrews (1988) holds for it. This result and B1 yield to the convergence

$$(NT_N)^{-1}\boldsymbol{\eta}^T\mathbf{B}_{NT_N}\boldsymbol{\eta} \xrightarrow{p} \boldsymbol{\eta}^T\mathbf{B}\boldsymbol{\eta}, \tag{A-3}$$

as $\left\{N, T_N\right\} \to \infty$, for any non-null $\boldsymbol{\eta} \in \mathbb{R}^m$. The existence of the limiting information matrix (22) follows the same methodology used in A.2.1 for the existence of (20), by considering B3′ instead of B3. The same notation $\mathbf{B}_N = (B_{k,l})_{k,l=1,\dots,m}$ and the same splits for each elements of the information matrix are adopted. So we highlight only the element which is different, i.e. $N^{-1}B_{12b} = N^{-1}\mathrm{E}(\mathbf{1}_N^T\mathbf{D}_t^{-1}\boldsymbol{\Sigma}_t\mathbf{D}_t^{-1}\mathbf{W}\mathbf{Y}_{t-1}) = N^{-1}B_{12a} + N^{-1}B_{12b}$. Clearly, $N^{-1}B_{12a} = \mu\mathbf{1}_N^T\boldsymbol{\Lambda}\mathbf{1}_N \to \mu f_1$. Moreover, when $i = j$, $\left|N^{-1}B_{12b}\right| = \left|N^{-1}\mathrm{E}\left[\sum_{i,j=1}^{N}\sigma_{ijt}(\mathbf{w}_i^T\tilde{\mathbf{Y}}_{t-1})/(\lambda_{i,t}\lambda_{j,t})\right]\right| = \left|N^{-1}H_{12b}\right| \to 0$, as $N \to \infty$. When $i \neq j$

$$\left|\frac{B_{12b}}{N}\right| \leq \frac{\beta_0^{-1}}{N}\mathrm{E}\left(\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|\rho_{ijt}\right|\sum_{i=1}^{N}\mathbf{w}_i^T\left|\tilde{\mathbf{Y}}_{t-1}\right|_v\right) \leq \frac{2\Phi\beta_0^{-1}}{N}\mathrm{E}\left(\mathbf{1}_N^T\mathbf{W}\left|\tilde{\mathbf{Y}}_{t-1}\right|_v\right),$$

which converges to 0, as $N \to \infty$, following (A-2). □

Proof of (2). Now we show asymptotic normality. Define $\varepsilon_{Nt} = \boldsymbol{\eta}^T\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}}^T\mathbf{D}_t^{-1}\boldsymbol{\xi}_t$, and recall the σ-field $\mathcal{F}_t^N = \sigma\left(\xi_{i,s} : 1 \leq i \leq N, s \leq t\right)$. 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. By $N^{-2}\mathrm{E}(\boldsymbol{\eta}^T\mathbf{s}_{Nt})^4 < \infty$, the Lindberg's condition is satisfied

$$\frac{1}{NT_N}\sum_{t=1}^{T_N}\mathrm{E}\left[\varepsilon_{Nt}^2 I\left(\left|\varepsilon_{Nt}\right| > \sqrt{NT_N}\delta\right) \mid \mathcal{F}_{t-1}^N\right] \leq \frac{\delta^{-2}}{N^2T_N^2}\sum_{t=1}^{T_N}\mathrm{E}\left(\varepsilon_{Nt}^4 \mid \mathcal{F}_{t-1}^N\right) \xrightarrow{p} 0,$$

for any $\delta > 0$, as $N \to \infty$. By the result in equation (A-3)

$$\frac{1}{NT_N}\sum_{t=1}^{T_N}\mathrm{E}\left(\varepsilon_{Nt}^2 \mid \mathcal{F}_{t-1}^N\right) = \frac{1}{NT_N}\sum_{t=1}^{T_N}\boldsymbol{\eta}^T\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}}^T\mathbf{D}_t^{-1}\mathrm{E}(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^T \mid \mathcal{F}_{t-1}^N)\mathbf{D}_t^{-1}\frac{\partial\boldsymbol{\lambda}_t}{\partial\boldsymbol{\theta}^T}\boldsymbol{\eta} \xrightarrow{p} \boldsymbol{\eta}^T\mathbf{B}\boldsymbol{\eta},$$

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, $(NT_N)^{-1/2}S_{NT_N} \xrightarrow{d} N(0, \boldsymbol{\eta}^T\mathbf{B}\boldsymbol{\eta})$, and an application of the Cramér-Wold theorem leads to the desired result. □

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12728