Submitted to Bernoulli
arXiv: arXiv:1704.02097

Multivariate Count Autoregression

KONSTANTINOS FOKIANOS[1,*] BÅRD STØVE[2,**] DAG TJØSTHEIM[2,†] and PAUL DOUKHAN[3,‡]

[1]*Department of Mathematics & Statistics, Lancaster University E-mail:* [*]k.fokianos@lancaster.ac.uk

[2]*Department of Mathematics, University of Bergen*
E-mail: [**]Bard.Stove@math.uib.no; [†]Dag.Tjostheim@math.uib.no

[3]*AGM, UMR 8088, University of Cergy-Pontoise and CIMFAV, Valparaiso E-mail:* [‡]doukhan@u-cergy.fr

We are studying linear and log-linear models for multivariate count time series data with Poisson marginals. For studying the properties of such processes we develop a novel conceptual framework which is based on copulas. Earlier contributions impose the copula on the joint distribution of the vector of counts by employing a continuous extension methodology. Instead we introduce a copula function on a vector of associated continuous random variables. This construction avoids conceptual difficulties related to the joint distribution of counts yet it keeps the properties of the Poisson process marginally. Furthermore, this construction can be employed for modeling multivariate count time series with other marginal count distributions. We employ Markov chain theory and the notion of weak dependence to study ergodicity and stationarity of the models we consider. Suitable estimating equations are suggested for estimating unknown model parameters. The large sample properties of the resulting estimators are studied in detail. The work concludes with some simulations and a real data example.

Keywords: autocorrelation, copula, ergodicity, generalized linear models, perturbation, prediction, stationarity, volatility.

1. Introduction

Modeling and inference of multivariate count time series is an important research topic; see [45] for a medical application, [47] for a financial application and more recently [49] for a marketing application and [39] for an environmental study. The interested reader is referred to the review paper by [33], for further details.

There are three main approaches taken towards the problem of modeling and inference for multivariate count time series. The first approach is based on the theory of integer autoregressive (INAR) models and was initiated by [25] and [35]. This work was further developed by [46, 47]. Estimation for INAR models is based on least squares methodology and likelihood based methods. However, even in the context of univariate INAR models, likelihood theory is quite cumbersome, especially for higher order models. Therefore, this class of models, which is adequate to describe some simple data structures, still poses challenges in terms of estimation (and prediction) especially when the model order is large.

The second class of models proposed for the analysis of count time series models, is that of parameter driven models. Recall that a parameter driven model (according to the broad categorization introduced by

[8]) is a model whose dynamics are driven by an unobserved process. In this case, state space models for multivariate count time series were studied by [31] and [32]; see also [48, 49], among others, for more recent contributions.

The aim of our contribution is to study models that fall within the class of observation driven models; that is models whose dynamics evolve according to past values of the process plus some noise. This is the case of the usual autoregressive models. In particular, observation driven models for count time series have been studied by [9], [21], [23] [10], among others. There is a growing recent literature in this topic; see [27], [38], [3], [1] and [36], for instance. These studies are concerned with linear count time series models. Although the linear model is adequate for several applications, it may not always be a natural candidate for count data analysis. In our view, log-linear models are more appropriate for general modeling of count time series. Some desirable properties of log-linear models include the ease of including covariates, incorporation of positive/negative correlation and avoiding parameter boundary problems; see [23], [2]. In fact, the log-linear model corresponds to the canonical link Poisson regression model for count data analysis; [41].

A major obstacle for the analysis of count time series is the choice of the joint count distribution. There are numerous proposals available in the literature generalizing the univariate Poisson probability mass function (pmf); some of these are reviewed in the previous references. However, the pmf of a multivariate Poisson discrete random vector is usually of quite complicated functional form and therefore maximum likelihood inference can be quite challenging (theoretically and numerically). Generally speaking, the choice of the joint distribution for multivariate count data is quite an interesting topic. In this work we address this problem by suggesting a copula based construction of a joint distribution. Instead of imposing a copula function on a vector of discrete random variables, we argue, based on Poisson process properties, that it can be introduced via a vector of continuous random variables. In this way, we avoid technical difficulties and we propose a plausible data generating process which keeps intact the properties of the Poisson properties, marginally. This approach can be extended to include other multivariate count distributions. Equipped with this construction and given a model, we suggest suitable estimating functions to estimate the unknown parameters. The main goals of this work are summarized by the following:

1. Develop a novel conceptual framework for studying count time series.
2. Give conditions for ergodicity and stationarity of both linear and log-linear models. The preferred methodologies are those of Markov chain theory (employing a perturbation approach) and theory of weak dependence. Although the linear model was treated by [38] in a parametric joint Poisson framework, we relax these conditions considerably when using the perturbation approach. For the log-linear model case, these conditions are new.
3. We suggest appropriate estimating functions which deliver consistent and asymptotically normally distributed estimators.

As a final remark we discuss the problem of proving stationarity and ergodicity of count time series. Following the discussions of [43] and [53, 54], the main difficulty is that the process itself consists of integer

valued random variables; however the mean process takes values on the positive real line and therefore it is quite challenging to prove ergodicity of the joint process (see also [4]). The study of theoretical properties of these models was initiated by the perturbation method suggested in [21] and was further developed in [43] (using the notion of β-mixing), [15] (weak dependence approach, see [16]), [56] and [13] (Markov chain theory without irreducibility assumptions) and [55] (based on the theory of e-chains; see [42]).

The paper is organized as follows: Section 2 discusses the basic modeling approach that we take towards modeling multivariate count time series. The copula structure which is imposed introduces dependence but without affecting the properties of the marginal Poisson processes. We will consider both a linear and a log-linear model. Section 3 gives the results about ergodic and stationary properties of the linear and log-linear models. Section 4 discusses Quasi Maximum Likelihood Estimation (QMLE) and shows that the resulting estimators are consistent and asymptotically normal. Section 5 presents a limited simulation study and a real data examples. The paper concludes with an appendix which contains the proofs of main results. Some further results are included in the supplementary material.

2. Model Assumptions

In what follows we assume that $\{\mathbf{Y}_t = (Y_{i,t}),\ i = 1, 2, \ldots, p,\ t = 1, 2 \ldots, \}$ denotes a p–dimensional count time series. Let $\{\boldsymbol{\lambda}_t = (\lambda_{i,t}),\ i = 1, 2, \ldots, p, t = 0, 1, \ldots\}$ be the corresponding p-dimensional intensity process and $\mathscr{F}_t^{\mathbf{Y},\boldsymbol{\lambda}}$ the σ–field generated by $\{\mathbf{Y}_0, \cdots, \mathbf{Y}_t, \boldsymbol{\lambda}_0\}$ with $\boldsymbol{\lambda}_0$ being a p-dimensional vector denoting the starting value of $\{\boldsymbol{\lambda}_t\}$. With this notation, the intensity process is given by $\boldsymbol{\lambda}_t = \mathrm{E}[\mathbf{Y}_t \mid \mathscr{F}_t^{\mathbf{Y},\boldsymbol{\lambda}}]$. We will be studying two autoregressive models for multivariate count time series analysis; the linear and log-linear models which are direct extensions of their univariate counterparts. The linear model is defined by assuming that for each $i = 1, 2, \ldots, p$,

$$Y_{i,t} \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}} \text{ is marginally Poisson}(\lambda_{i,t}),\ \ \boldsymbol{\lambda}_t = \mathbf{d} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1}, \tag{1}$$

where $\mathbf{d}$ is a p-dimensional vector and $\mathbf{A}$, $\mathbf{B}$ are $p \times p$ unknown matrices. *The elements of* $\mathbf{d}$, $\mathbf{A}$ *and* $\mathbf{B}$ *are assumed to be positive such that* $\lambda_{i,t} > 0$. *for all* i *and* t. Model (1) generalizes naturally the linear autoregressive model discussed by [50], [19] and [21], among others. The log-linear model that we consider is the multivariate analogue of the univariate log-linear model proposed by [23]. More precisely assume that for each $i = 1, 2, \ldots, p$,

$$Y_{i,t} \mid \mathscr{F}_t^{\mathbf{Y},\boldsymbol{\lambda}} \text{ is marginally Poisson}(\lambda_{i,t}),\ \ \boldsymbol{\nu}_t = \mathbf{d} + \mathbf{A}\boldsymbol{\nu}_{t-1} + \mathbf{B}\log(\mathbf{Y}_{t-1} + \mathbf{1}_p), \tag{2}$$

where $\boldsymbol{\nu}_t \equiv \log \boldsymbol{\lambda}_t$ is defined componentwise (i.e. $\nu_{i,t} = \log \lambda_{i,t}$) and $\mathbf{1}_p$ denotes the p–dimensional vector which consists of ones. *In the case of* (2), *we do not impose any positivity constraints on the parameters* $\mathbf{d}$, $\mathbf{A}$ *and* $\mathbf{B}$; *this is an important argument favoring the log-linear model.* The log-linear model (2) is expected to be a better candidate for count data observed jointly with some other covariate time series or where negative correlation is observed. Indeed, if $\mathbf{X}_t$ is a covariate vector of dimension p, then the second equation of (2)

can be rewritten as $\boldsymbol{\nu}_t = \mathbf{d} + \mathbf{A}\boldsymbol{\nu}_{t-1} + \mathbf{B}\log(\mathbf{Y}_{t-1} + \mathbf{1}_p) + \mathbf{C}\mathbf{X}_t$ for a $p \times p$ matrix $\mathbf{C}$. In addition, we show in Sec. **??** of the supplement that the model induces both positive and negative correlation.

A fundamental problem in the analysis of multivariate count data is the specification of a joint distribution for the counts. There are numerous proposals made in the literature aiming at generalizing the univariate Poisson assumption to the multivariate case but the resulting joint distributions are quite complex for likelihood based inference. A possible construction can be based on independent Poisson random variables or on copulas and mixture models (see [30, Ch. 37], [29, Sec 7.2]). However, the resulting joint pmf is complicated and therefore the log-likelihood function cannot be calculated analytically (or, sometimes, even approximated). We propose a different approach. Consider the first equation of (1) but the same discussion applies to (2) subject to minor modifications. It implies that each component $Y_{i,t}$ is *marginally* a Poisson process. But the joint distribution of the vector $\{\mathbf{Y}_t\}$ is not necessarily distributed as a multivariate Poisson random variable. Our general construction, as outlined below, allows for arbitrary dependence among the marginal Poisson components by utilizing fundamental properties of the Poisson process. We give a detailed account of the data generating process. Suppose that $\boldsymbol{\lambda}_0 = (\lambda_{1,0}, \ldots, \lambda_{p,0})$ is some starting value. Then consider the following data generating mechanism:

1. Let $\mathbf{U}_l = (U_{1,l}, \ldots, U_{p,l})$ for $l = 1, 2, \ldots, K$, be a sample from a p-dimensional copula $C(u_1, \ldots, u_p)$. Then $U_{i,l}$, $l = 1, 2, \ldots, K$ follow marginally the uniform distribution on $(0, 1)$, for $i = 1, 2, \ldots, p$.
2. Consider the transformation $X_{i,l} = -\log U_{i,l}/\lambda_{i,0}, \quad i = 1, 2, \ldots, p$. Then, the marginal distribution of $X_{i,l}$, $l = 1, 2, \ldots, K$ is exponential with parameter $\lambda_{i,0}$, $i = 1, 2, \ldots, p$.
3. Define now (taking K large enough) $Y_{i,0} = \max_{1 \le k \le K} \left\{ \sum_{l=1}^{k} X_{i,l} \le 1 \right\}, \quad i = 1, 2, \ldots, p$. Then $\mathbf{Y}_0 = (Y_{1,0}, \ldots, Y_{p,0})$ is marginally a set of first values of a Poisson process with parameter $\boldsymbol{\lambda}_0$.
4. Use model (1) (respectively (2)) to obtain $\boldsymbol{\lambda}_1$.
5. Return back to step 1 to obtain $\mathbf{Y}_1$, and so on.

The aforementioned construction of the joint distribution of the counts imposes the dependence among the components of the vector process $\{\mathbf{Y}_t\}$ by taking advantage of a *copula structure on the waiting times of the Poisson process*. Equivalently, the copula is imposed on the uniform random variables generating the exponential waiting times. Such an approach does not pose any problems on obtaining the joint distribution of the random vector $\{\mathbf{Y}_t\}$ which is composed of discrete valued random variables. This can be extended to other marginal count processes if they can be generated by continuous inter arrival times. For instance, suppose that $Y_{i,t}$ is marginally mixed Poisson with mean $Z_{i,t}\lambda_{i,t}$ where $Z_{i,t}$ is an iid sequence for all $i = 1, 2, \ldots, p$, it is independent of $\mathbf{Y}_t$ for all t and satisfies $\mathrm{E}[Z_{i,t}] = 1$ (see [7]). Many families of count distributions, including the negative binomial, can be generated by this construction. Then steps 1-5 of the above algorithm still can be used to generate data from a count time series models whose marginals are not necessarily Poisson. Indeed, generating at the first step an additional vector $Z_{i,0}$, say $z_{i,0}$ define again at step 2 the waiting times by $X_{i,l} = -\log U_{i,l}/z_{i,0}\lambda_{i,0}, \quad i = 1, 2, \ldots, p$. Then, the distribution of $X_{i,0}$ is mixed exponential and therefore steps 3-5 deliver a realization of a count vector whose marginal distribution is mixed Poisson.

An added advantage of this approach is that copula is defined uniquely for continuous multivariate random variables. For a lucid discussion about copula for discrete multivariate distributions, see [26], in particular pp. 507-508. Our approach is different from the approach taken by [27]. These authors replace the original counts by employing the continued extension method of [12]. Accordingly, they add some noise of the form $U-1$, where U is uniform, to counts to transform them to continuous random variables such that the problem of copula identifiability is bypassed. This is an interesting idea. Under an assumption of small dispersion asymptotics a covariance structure is obtained which is similar to that obtained for the linear model in Sec. **??** of the supplement. Note that the continued extension method of [27] has been investigated in a simulation study by [44]. In our approach, there is need to distinguish between the copula on the counts themselves and the copula on the waiting times. The transformation from waiting times to counts is stochastic, and while the copula as such is invariant to one-to-one deterministic transformations, we do not have such a transformation in our case. Hence, the instantaneous correlation among the components of vector of counts is not equal to the correlation induced by the copula imposed to the vector of waiting times. Therefore, the interpretation of the instantaneous correlation for both linear and log-linear models is associated with the correlation of the vector of waiting times and should be done with care. An initial approach of estimating the correlation among waiting times is discussed immediately after Thm. 4.2.

Hence, the first equation of model (1) can be restated as

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \mathbf{d} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1} \tag{3}$$

where $\{\mathbf{N}_t\}$ is a sequence of independent p-variate copula–Poisson processes which counts the number of events in $[0, \lambda_{1,t}] \times \ldots \times [0, \lambda_{p,t}]$. We also define the multivariate log–linear model (2) by

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\nu}_t), \quad \boldsymbol{\nu}_t = \mathbf{d} + \mathbf{A}\boldsymbol{\nu}_{t-1} + \mathbf{B}\log(\mathbf{Y}_{t-1} + \mathbf{1}_p) \tag{4}$$

Now, the process $\{\mathbf{N}_t\}$ denotes as before a sequence of independent p-variate copula–Poisson processes which counts the number of events in $[0, \exp(\nu_{1,t})] \times \ldots \times [0, \exp(\nu_{p,t})]$. In the supplement (Sec. **??**) we derive the theoretical autocovariance matrices of models (3) (see also [27]) and we show that all their elements are positive and depend on the joint distribution of the count vector which in turn depends on the copula structure. The positivity of all elements shows that linear models can be applied to time series like the one we consider in Section 5; see Fig. 2. In addition, we derive, approximately, the autocovariance function of $\mathbf{W}_t \equiv \log(\mathbf{Y}_t + \mathbf{1}_p)$ for model (4). Its form shows that we can have both positive and negative correlation. Explicit calculation of the autocovariance function of $\mathbf{Y}_t$ for (4) is a challenging problem which can be studied by simulation.

It is instructive to consider model (3) in more detail because its structure is closely related to the theory of GARCH models, [6]. Observe that each component of the vector-process $\{\mathbf{Y}_t\}$ is distributed as a Poisson random variable. But the mean of a Poisson random variable equals its variance; therefore model (3) resembles some structure of multivariate GARCH model, see [40] and [24]. For $p = 2$, for example, the second

equation of (3) becomes

$$\begin{aligned}\lambda_{1,t} &= d_1 + a_{11}\lambda_{1,t-1} + a_{12}\lambda_{2,t-1} + b_{11}Y_{1,t-1} + b_{12}Y_{2,t-1},\\ \lambda_{2,t} &= d_2 + a_{21}\lambda_{1,t-1} + a_{22}\lambda_{2,t-1} + b_{21}Y_{1,t-1} + b_{22}Y_{2,t-1},\end{aligned}$$

where d_i is the ith element of $\mathbf{d}$ and a_{ij} (b_{ij}, respectively) is the (i,j)th element of $\mathbf{A}$ ($\mathbf{B}$, respectively). We can give the following interpretation to model parameters. When $a_{12} = b_{12} = 0$, then λ_{1t} depends only on its own past. If this is not true, then the parameters denote the linear dependence of λ_{1t} on $\lambda_{2,t-1}$ and $Y_{2,t-1}$ in the presence of $\lambda_{1,t-1}$ and $Y_{1,t-1}$. Similar results hold when $a_{21} = b_{21} = 0$ and the previous discussion applies to the case of (4).

3. Ergodicity and Stationarity

Towards the analysis of models (3) and (4), we employ the perturbation techniques as developed by [21] and [23]. In addition, we include a study which is based on the notion of weak dependence (for more, see [16] and [11]). Both approaches are employed and compared for obtaining ergodicity and stationarity of (3) and (4). In fact, the main goal is to obtain stationarity and ergodicity of the joint process $(\mathbf{Y}_t, \boldsymbol{\lambda}_t)$. For the specific examples of processes given by (3) and (4) the sufficient conditions obtained by the perturbation and weak dependence approach are different; however all proofs are based on a contraction property of the process $\{\boldsymbol{\lambda}_t\}$ (in the case of (3)) and $\{\boldsymbol{\nu}_t\}$ (in the case of (4)). Note that the copula construction is not used in the proof of ergodicity by neither of the approaches we take nor it is used in the estimation of the parameters. It is only used in the proofs of Lemmas 3.1-3.2 (with no additional conditions, however) to show that the perturbed models is close to non-perturbed models via the Markov chain approach we take. In this respect the situation may be similar to a multivariate ARMA model where the stability conditions are independent of the correlations in the innovations. Similar comments can be made about the multivariate GARCH. The correlation structure may not necessarily be used in the estimation of the parameter matrices for these processes either, but this may lead to estimators that are not efficient. Whereas use of correlation in the innovations does not lead to an extension of ARMA or GARCH, staying within the multivariate Poisson is troublesome because of the very complicated and restricting nature of this model. It is then natural to allow for a more general dependence structure between Poisson components, and the copula seems to be a natural instrument for describing such dependence, which leads to a quite flexible model. The copula modeling of dependence is explicitly used in Section 5 of the paper just to produce such flexible models.

We denote by $\|\mathbf{x}\|_d = (\sum_{i=1}^p |x_i|^d)^{1/d}$ the l^d- norm of a p-dimensional vector $\mathbf{x}$. For a $q \times p$ matrix $\mathbf{A} = (a_{ij})$, $i = 1, \ldots, q, j = 1, \ldots, p$, we let $|||\mathbf{A}|||_d$ denote the generalized matrix norm $|||\mathbf{A}|||_d = \max_{\|\mathbf{x}\|_d=1} \|\mathbf{A}\mathbf{x}\|_d$. If $d = 1$, then $|||\mathbf{A}|||_1 = \max_{1\le j\le p} \sum_{i=1}^q |a_{ij}|$, and when $d = 2$, $|||\mathbf{A}|||_2 = \rho^{1/2}(\mathbf{A}^T\mathbf{A})$ where $\rho(.)$ denotes the spectral radius The Frobenius norm is denoted by $|||\mathbf{A}|||_F = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}$. If $q = p$, then these norms are matrix norms.

3.1. Linear Model

Following [21], we introduce the perturbed model

$$\mathbf{Y}_t^m = \mathbf{N}_t(\boldsymbol{\lambda}_t^m), \;\; \boldsymbol{\lambda}_t^m = \mathbf{d} + \mathbf{A}\boldsymbol{\lambda}_{t-1}^m + \mathbf{B}\mathbf{Y}_{t-1}^m + \boldsymbol{\epsilon}_t^m, \tag{5}$$

where $\boldsymbol{\epsilon}_t^m = c_m \mathbf{V}_t$. Here the sequence c_m is strictly positive and tends to zero, as $m \to \infty$, and $\mathbf{V}_t$ is a p-dimensional vector which consists of independent positive random variables each of which having a bounded support of the form $[0, M]$, for some $M > 0$. The introduction of the perturbed process allows to prove ergodicity and stationarity of the joint process $\{(\mathbf{Y}_t^m, \boldsymbol{\lambda}_t^m, \boldsymbol{\epsilon}_t^m)\}$. The first result is given by the following proposition:

Proposition 3.1. Consider model (5) and suppose that $|||\mathbf{A} + \mathbf{B}|||_2 < 1$. Then the process $\{\boldsymbol{\lambda}_t^m, t > 0\}$ is a geometrically ergodic Markov chain with finite r'th moments, for any $r > 0$. Moreover, the process $\{(\mathbf{Y}_t^m, \boldsymbol{\lambda}_t^m, \boldsymbol{\epsilon}_t), t > 0\}$ is $V_{\mathbf{Y},\boldsymbol{\lambda},\boldsymbol{\epsilon}}$ geometrically ergodic Markov chain with $V_{\mathbf{Y},\boldsymbol{\lambda},\epsilon} = 1 + \|\mathbf{Y}\|_2^r + \|\boldsymbol{\lambda}\|_2^r + \|\boldsymbol{\epsilon}\|_2^r$, $r > 0$.

The following results show that as $c_m \to 0$ as $m \to \infty$, then the difference between (3) and (5) can be made arbitrary small.

Lemma 3.1. Consider models (3) and (5). If $|||\mathbf{A} + \mathbf{B}|||_2 < 1$, then the following hold true:

1. $\|\mathrm{E}(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t)\|_2 = \|\mathrm{E}(\mathbf{Y}_t^m - \mathbf{Y}_t)\|_2 \leq \delta_{1,m}$.
2. $\mathrm{E}\|(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t)\|_2^2 \leq \delta_{2,m}$.
3. $\mathrm{E}\|(\mathbf{Y}_t^m - \mathbf{Y}_t)\|_2^2 \leq \delta_{3,m}$.

In the above $\delta_{i,m} \to 0$, as $m \to \infty$. In addition, for sufficiently large m, $\|\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t\|_2 \leq \delta$ and $\|\mathbf{Y}_t^m - \mathbf{Y}_t\|_2 \leq \delta$, almost surely, for any $\delta > 0$.

The above results show that the condition $|||\mathbf{A}+\mathbf{B}|||_2 < 1$ is sufficient to guarantee the required contraction (c.f. Lemma (3.1)) and existence of all moments of the joint process $\{(\mathbf{Y}_t, \boldsymbol{\lambda}_t)\}$, (see Proposition (3.1)). In the simple case of a vector autoregressive model with $\mathbf{A} = \mathbf{0}$ in (3), the condition $|||\mathbf{B}|||_2 < 1$ guarantees stationarity and ergodicity of the process $\{\mathbf{Y_t}\}$. This fact is proved by iterating the recursions of the autoregressive model yielding powers of $\mathbf{B}$. However, this technique cannot be applied to the general multivariate case but it is deduced by Proposition 3.1. We conjecture that for the general linear multivariate model of order (q, ℓ)

$$\boldsymbol{\lambda}_t = \mathbf{d} + \sum_{i=1}^{\ell} \mathbf{A}_i \boldsymbol{\lambda}_{t-i} + \sum_{j=1}^{q} \mathbf{B}_j \mathbf{Y}_{t-j},$$

the condition $\sum_{i=1}^{\max(\ell,q)} |||\mathbf{A}_i + \mathbf{B}_i|||_2 < 1$ is sufficient for proving Proposition 3.1.

We turn now to an alternative method; namely we will use the concept of weak dependence to study the properties of the linear model (3). This approach does not require a perturbation argument but the sufficient conditions obtained are weaker. The proof of this result parallels the proof of [15]; we outline some aspects of it in the appendix.

Proposition 3.2. Consider model (3) and suppose that $|||\mathbf{A}|||_1 + |||\mathbf{B}|||_1 < 1$. Then there exists a unique causal solution $\{(\mathbf{Y}_t, \boldsymbol{\lambda}_t)\}$ to model (3) which is stationary, ergodic and satisfies $\mathrm{E}\|\mathbf{Y}_t\|_r^r < \infty$ and $\mathrm{E}\|\boldsymbol{\lambda}_t\|_r^r < \infty$, for any $r \in \mathbb{N}$.

The closest result reported in the literature analogous to those obtained by Propositions 3.1 and 3.2 can be found in [38, Prop. 4.2.1] which is, in fact, based on the assumption of a joint multivariate Poisson distribution for the vector of counts. The author shows that if there exists a $p \geq 1$ such that $|||\mathbf{A}|||_p + 2^{1-(1/p)}|||\mathbf{B}|||_p < 1$ then the process $\{\boldsymbol{\lambda}_t\}$ is geometrically moment contracting, see [57] for definition. In the case that $p = 2$, then the condition of Proposition 3.1 improves this result for the perturbed process $\{\boldsymbol{\lambda}_t^m\}$. When $p = 1$ we see that the aforementioned condition is reduced to that proved in Proposition 3.2. As a closing remark, note that (3) can be iterated to obtain

$$\boldsymbol{\lambda}_t = \sum_{j=0}^{k-1} \mathbf{A}^j \mathbf{d} + \mathbf{A}^k \boldsymbol{\lambda}_{t-k} + \sum_{j=0}^{k-1} \mathbf{A}^j \mathbf{B} \mathbf{Y}_{t-j-1}. \tag{6}$$

for $k \in \mathbb{N}$. Assume that $|||\mathbf{A}|||_2 < 1$. Then an alternative representation of model (1) holds, from a passage to the limit, as $k \uparrow \infty$, from the above equation:

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \qquad \boldsymbol{\lambda}_t = (\mathbf{I}_p - \mathbf{A})^{-1}\mathbf{d} + \sum_{j=0}^{\infty} \mathbf{A}^j \mathbf{B} \mathbf{Y}_{t-j-1}. \tag{7}$$

where $\mathbf{I}_p$ is the identity matrix of order p In this case, the stationarity condition obtained from [17], as a multivariate variant of [15], is given by

$$\sum_{j=0}^{\infty} |||\mathbf{A}^j \mathbf{B}|||_2 < 1. \tag{8}$$

This condition is implied from $|||\mathbf{A}|||_2 + |||\mathbf{B}|||_2 < 1$. Indeed, $|||\mathbf{A}^j\mathbf{B}|||_2 \leq |||\mathbf{A}|||_2^j \cdot |||\mathbf{B}|||_2$ and therefore

$$\sum_{j=0}^{\infty} |||\mathbf{A}^j \mathbf{B}|||_2 \leq \sum_{j=0}^{\infty} |||\mathbf{A}|||^j \cdot |||\mathbf{B}|||_2 = \frac{|||\mathbf{B}|||_2}{1 - |||\mathbf{A}_2|||} < 1.$$

In other words, (8) improves Proposition 3.2. However, if $\mathbf{AB} = \mathbf{BA}$ and if they are non-negative definite, then we obtain that $|||\mathbf{A} + \mathbf{B}|||_2 = |||\mathbf{A}|||_2 + |||\mathbf{B}|||_2$ and then all obtained conditions coincide. To see that holds true, note that when $\mathbf{AB} = \mathbf{BA}$ then $\mathbf{A}, \mathbf{B}$ can be simultaneously reduced in triangular blocks with the same eigenvalue on each block.

3.2. Log-linear Model

We turn to the study of the log–linear model (4). We introduce again its perturbed version by

$$\mathbf{Y}_t^m = \mathbf{N}_t(\boldsymbol{\nu}_t^m), \;\; \boldsymbol{\nu}_t^m = \mathbf{d} + \mathbf{A}\boldsymbol{\nu}_{t-1}^m + \mathbf{B}\log(\mathbf{Y}_{t-1}^m + \mathbf{1}_p) + \boldsymbol{\epsilon}_t^m, \tag{9}$$

where the perturbation has the same structure as in (5); . Then, [23, Lemma A.2] show that $\mathrm{E}[(\log(Y_{j,t-1}^m + 1))^r | \nu_{j;t-1} = \nu_j] \sim \nu_j^r$, $j = 1, 2, \ldots, p$ and $r > 0$. Therefore, we can employ similar arguments as those employed in [23] to prove the following results.

Proposition 3.3. Consider (9) and suppose that $|||\mathbf{A}|||_2 + |||\mathbf{B}|||_2 < 1$. Then the process $\{\boldsymbol{\nu}_t^m, t > 0\}$ is geometrically ergodic Markov chain with finite r'th moments, for any $r > 0$. Moreover, the process $\{(\mathbf{Y}_t^m, \boldsymbol{\nu}_t^m, \boldsymbol{\epsilon}_t), t > 0\}$ is $V_{\mathbf{Y},\boldsymbol{\nu},\boldsymbol{\epsilon}}$ geometrically ergodic Markov chain with $V_{\mathbf{Y},\boldsymbol{\lambda},\boldsymbol{\epsilon}} = 1 + \|\log(\mathbf{Y} + \mathbf{1}_p)\|_2^{2r} + \|\boldsymbol{\nu}\|_2^{2r} + \|\boldsymbol{\epsilon}\|_2^{2r}$, $r > 0$.

The proof of the above result is omitted. However, we give in the appendix some details about the following approximation lemma.

Lemma 3.2. Consider models (4) and (9). If $|||\mathbf{A}|||_2 + |||\mathbf{B}|||_2 < 1$, then the following hold true:

1. $\|\mathrm{E}(\boldsymbol{\nu}_t^m - \boldsymbol{\nu}_t\|_2 \to 0$, as $m \to \infty$ and $\|\mathrm{E}(\mathbf{Y}_t^m - \mathbf{Y}_t)\|_2 \leq \delta_{1,m}$.
2. $\mathrm{E}\|(\boldsymbol{\nu}_t^m - \boldsymbol{\nu}_t)\|_2^2 \leq \delta_{2,m}$.
3. $\mathrm{E}\|(\mathbf{Y}_t^m - \mathbf{Y}_t)\|_2^2 \leq \delta_{3,m}$.
4. $\mathrm{E}\|(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t)\|_2^2 \leq \delta_{4,m}$.

In the above $\delta_{i,m} \to 0$, as $m \to \infty$. In addition, for sufficiently large m, $\|\boldsymbol{\nu}_t^m - \boldsymbol{\nu}_t\|_2 \leq \delta$ and $\|\mathbf{Y}_t^m - \mathbf{Y}_t\|_2 \leq \delta$, almost surely, for any $\delta > 0$.

We see that the condition $|||\mathbf{A} + \mathbf{B}|||_2 < 1$ obtained for the linear model (3) is not implied by the condition $|||\mathbf{A}|||_2 + |||\mathbf{B}|||_2 < 1$ which was found for the log-linear model. Recall that in the case of the linear model (3) all parameters are assumed to be positive for ensuring that the components of $\boldsymbol{\lambda}_t$ are positive. This is not necessary for the log-linear model case. Closing this section, we note that the weak dependence approach delivers a similar condition.

Proposition 3.4. Consider model (4) and suppose that $|||\mathbf{A}|||_1 + |||\mathbf{B}|||_1 < 1$. Then there exists a unique causal solution $\{(\mathbf{Y}_t, \boldsymbol{\nu}_t)\}$ to model (2) which is stationary, ergodic and satisfies $\mathrm{E}\|\log(\mathbf{Y}_t + \mathbf{1}_p)\|_r^r < \infty$ and $\mathrm{E}\|\boldsymbol{\nu}_t\|_r^r < \infty$ and $\mathrm{E}[\exp(r\|\boldsymbol{\nu}_t\|_1)] < \infty$ for any $r \in \mathbb{N}$.

The same remarks made for the linear model (3) in page 8 hold true for the case of the log-linear model (4). Indeed, note that the infinite representation is still valid by replacing $\boldsymbol{\lambda}_t$ by $\boldsymbol{\nu}_t$ and $\mathbf{Y}_t$ by $\log(\mathbf{Y}_t + \mathbf{1}_p)$. Hence, (8) asserts stationarity and weak dependence for the log-linear model. In both cases we were not able to prove the conjecture that $|||\mathbf{A} + \mathbf{B}|||_2 < 1$ implies weak dependence. However, (8) improves on the results of Lemmas 3.2 and 3.4.

4. Quasi-Likelihood Inference

Suppose that $\{\mathbf{Y}_t, t = 1, 2, \ldots, n\}$ is an available sample from a count time series and denote the vector of unknown parameters by $\boldsymbol{\theta}$; that is $\boldsymbol{\theta}^T = (\mathbf{d}^T, \mathrm{vec}^T(\mathbf{A}), \mathrm{vec}^T(\mathbf{B}))$, where $\mathrm{vec}(\cdot)$ denote the vec operator and $\dim(\theta) \equiv d = p(1 + 2p)$. The general approach that we take towards the estimation problem is based on the theory of estimating functions as outlined by [37] for longitudinal data analysis and [5], [28], among others, for stochastic processes. We will be considering the following conditional quasi–likelihood function, given $\boldsymbol{\lambda}_0$, for the parameter vector $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n}\prod_{i=1}^{p}\Big\{\frac{\exp(-\lambda_{i,t}(\boldsymbol{\theta}))\lambda_{i,t}^{y_{i,t}}(\boldsymbol{\theta})}{y_{i,t}!}\Big\}.$$

This is equivalent to considering model (1) (and (2)) under the assumption of contemporaneous independence among time series. This assumption simplifies computation of estimators and their respective standard errors. At the same time, it guarantees consistency and asymptotic normality of the resulting estimator (see [7], [2] and [14] for recent contributions in the context of count time series). The main idea is bsed on the correct mean model specification. In other words, if we assume that for a given count time series and regardless of the true data generating process, there exists a "true" vector of parameters, say $\boldsymbol{\theta}_0$, such that (1) holds (respectively (2)), then we obtain consistent and asymptotically normally distributed estimators by maximizing the quasi log-likelihood function (10). This result carries over to the Double Exponential model considered by [27] but is should be applied with some care because [18] has shown that the conditional expectation of this distribution is approximately λ_t. **We are not aware of any results relating the Double Poisson distribution to properties of Poisson type processes, so Prop. 3.1 and 3.3 are not applicable to this class of models.** We point out that [1], independent of us, considered the same approach but his work neither gives conditions for ergodicity for the models we examine nor does it consider log-linear multivariate models. In the following, we give some details for the linear model case but inference can be easily developed for the log–linear model (2) following the same arguments; we will only highlight some different aspects of each model.

The quasi log-likelihood function is equal to

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{p}\Big(y_{i,t}\log\lambda_{i,t}(\boldsymbol{\theta}) - \lambda_{i,t}(\boldsymbol{\theta})\Big). \tag{10}$$

We denote by $\widehat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$, the QMLE of $\boldsymbol{\theta}$. The score function is given by

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{p}\Big(\frac{y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta})} - 1\Big)\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{t=1}^{n}\frac{\partial\boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{D}_t^{-1}(\boldsymbol{\theta})\Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta})\Big) \equiv \sum_{t=1}^{n}s_t(\boldsymbol{\theta}), \tag{11}$$

where $\partial\boldsymbol{\lambda}_t/\partial\boldsymbol{\theta}^T$ is a $p \times d$ matrix and $\mathbf{D}_t$ is the $p \times p$ diagonal matrix with the i'th diagonal element equal to $\lambda_{i,t}(\theta)$, $i = 1, 2, \ldots, p$. Straightforward differentiation shows that under model (1), we obtain the following

recursions:

$$\begin{aligned}\frac{\partial \boldsymbol{\lambda}_t}{\partial \mathbf{d}^T} &= \mathbf{I}_p + \mathbf{A}\frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \mathbf{d}^T},\\ \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{A})} &= (\boldsymbol{\lambda}_{t-1}\otimes \mathbf{I}_p)^T + \mathbf{A}\frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \operatorname{vec}^T(\mathbf{A})},\\ \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{B})} &= (\mathbf{Y}_{t-1}\otimes \mathbf{I}_p)^T + \mathbf{A}\frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \operatorname{vec}^T(\mathbf{B})},\end{aligned} \tag{12}$$

where $\otimes$ denotes Kronecker's product. The Hessian matrix is given by

$$\mathbf{H}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{p}\frac{y_{i;t}}{\lambda_{i,t}^2(\boldsymbol{\theta})}\frac{\partial \lambda_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \lambda_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} - \sum_{t=1}^{n}\sum_{i=1}^{p}\Big(\frac{y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta})}-1\Big)\frac{\partial^2 \lambda_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}. \tag{13}$$

Therefore, the conditional information matrix is equal to

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\mathbf{D}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}_t(\boldsymbol{\theta})\mathbf{D}_t^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}, \tag{14}$$

where the matrix $\boldsymbol{\Sigma}_t(\cdot)$ denotes the *true* covariance matrix of the vector $\mathbf{Y}_t$. In case that the process $\{\mathbf{Y}_t\}$ consists of uncorrelated components then $\boldsymbol{\Sigma}_t(\boldsymbol{\theta}) = \mathbf{D}_t(\boldsymbol{\theta})$. We will study the asymptotic properties of the QMLE $\widehat{\boldsymbol{\theta}}$. By using [52, Thm 3.2.23] which is based on the work by [34], we can prove existence, consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}$. Continuous differentiability of the log-likelihood function, which is guaranteed by the Poisson assumption, is instrumental for obtaining these results. The main problem that we are faced with is that we cannot use directly the sufficient ergodicity and stationarity conditions for the unperturbed model to obtain the asymptotic theory (see also [21], [23] and [53, 54] for detailed discussion about the issues involved). Therefore we use the corresponding conditions for the perturbed model and then show that the perturbed and unperturbed versions are "close". Towards this goal define analogously S_n^m to be the MQLE score function for the perturbed model with $(\mathbf{Y}_t, \boldsymbol{\lambda}_t)$ replaced by $(\mathbf{Y}_t^m, \boldsymbol{\lambda}_t^m)$. Then, Theorem 4.1 follows immediately after proving Lemmas 4.1-4.3 and taking into account Remark 4.1 concerning the third derivative of the log-likelihood function. Together these results verify the conditions of [52, Thm 3.2.23]. Lemma 4.1 is proved in the appendix while Lemmas 4.2 and 4.3 are proved in the supplement.

Lemma 4.1. Define the matrices (see (15))

$$\mathbf{G}^m(\boldsymbol{\theta}) = \mathrm{E}\Big(s_t^m(\boldsymbol{\theta})s_t^m(\boldsymbol{\theta})^T\Big) \quad \text{and} \quad \mathbf{G}(\boldsymbol{\theta}) = \mathrm{E}\Big(s_t(\boldsymbol{\theta})s_t(\boldsymbol{\theta})^T\Big).$$

Under the assumptions of Theorem 4.1 the above matrices evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, satisfy $\mathbf{G}^m \to \mathbf{G}$, as $m \to \infty$.

Lemma 4.2. Under the assumptions of Theorem 4.1 the score functions for the perturbed (5) and unperturbed model (4) evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ satisfy the following:

1. $S_n^m/n \xrightarrow{\text{a.s}} 0$,
2. $S_n^m/\sqrt{n} \xrightarrow{\text{d}} S^m := N(0, \mathbf{G}^m)$,
3. $S^m \xrightarrow{\text{d}} N(0, \mathbf{G})$, as $m \to \infty$,
4. $\lim_{m\to\infty} \limsup_{n\to\infty} P(||S_n^m - S_n||_2 > \epsilon\sqrt{n}) = 0, \quad \forall\ \epsilon > 0.$

Lemma 4.3. Recall the Hessian matrix defined by (13), $\mathbf{H}_n$, and let $\mathbf{H}_n^m$ be the Hessian matrix which corresponds to the perturbed model (5) evaluated at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, under the assumptions of Theorem 4.1

1. $\mathbf{H}_n^m \xrightarrow{\text{P}} \mathbf{H}^m$ as $n \to \infty$
2. $\lim_{m\to\infty} \limsup_{n\to\infty} P(|||\mathbf{H}_n^m - \mathbf{H}_n|||_2 > \epsilon n) = 0, \quad \forall \epsilon > 0.$

where $\mathbf{H}$ is given by (16) (and analogously for $\mathbf{H}^m$). In addition, the matrix $\mathbf{H}$ is positive definite.

Theorem 4.1. Consider model (3). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. Suppose that $\boldsymbol{\Theta}$ is compact and assume that the true value $\boldsymbol{\theta}_0$ belongs to the interior of $\boldsymbol{\Theta}$. Suppose that at the true value $\boldsymbol{\theta}_0$, the condition of Proposition 3.1 hold true. Then there exists a fixed open neighborhood, say $O(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} :\ \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 < \delta\}$, of $\boldsymbol{\theta}_0$ such that with probability tending to 1 as $n \to \infty$, the equation $S_n(\boldsymbol{\theta}) = 0$ has a unique solution, say $\widehat{\boldsymbol{\theta}}$. Furthermore, $\widehat{\boldsymbol{\theta}}$ is strongly consistent and asymptotically normal,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\text{d}} N(0, \mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1})$$

where the matrices $\mathbf{G}(\boldsymbol{\theta})$ and $\mathbf{H}(\boldsymbol{\theta})$ are defined by

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

$$\mathbf{H}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right] \tag{16}$$

and expectation is taken with respect to the stationary distribution of $\{\mathbf{Y}_t\}$.

When the components of the time series $\{\mathbf{Y}_t\}$ are uncorrelated, then $\boldsymbol{\Sigma}_t = \mathbf{D}_t$ and therefore the matrices $\mathbf{G}$ and $\mathbf{H}$ coincide. Hence, we obtain a standard result for the ordinary MLE in this case. All the above quantities can be calculated by their respective sample counterparts.

Remark 4.1. To conclude the proof of Theorem 4.1 we need to show that the expected value of all third derivatives of the log-likelihood function (10) of the perturbed model (5) within the neighborhood of the true parameter $O(\boldsymbol{\theta}_0)$ are uniformly bounded. Additionally, we need to show that the all third derivatives of the unperturbed model (3) are "close" to the third derivatives of (5). This point was documented in several publications including [21] (for the case of linear model) and [23] (for the case of the log-linear model). In the supplement, we outline the methodology of obtaining this result.

We consider briefly QMLE inference for the case of the log-linear model (4). Given the log-likelihood function (10) we obtain the score, Hessian matrix and conditional information matrix by

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{p}\Big(y_{i,t} - \exp(\nu_{i,t}(\boldsymbol{\theta}))\Big)\frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n}\frac{\partial \boldsymbol{\nu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big(\mathbf{Y}_t - \exp(\boldsymbol{\nu}_t(\boldsymbol{\theta})\Big), \tag{17}$$

$$\mathbf{H}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{p}\exp(\nu_{i,t}(\boldsymbol{\theta}))\frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} - \sum_{t=1}^{n}\sum_{i=1}^{p}\Big(y_{i,t} - \exp(\nu_{i,t}(\boldsymbol{\theta}))\Big)\frac{\partial^2 \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T},$$

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{p}\exp(\nu_{i,t}(\boldsymbol{\theta}))\frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \nu_{i,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T},$$

respectively. The recursions for $\partial \boldsymbol{\nu}_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$ required for computing the QMLE are obtained as in (12) but with $\boldsymbol{\lambda}_t$ replaced by $\boldsymbol{\nu}_t$ and $\mathbf{Y}_{t-1}$ by $\log(\mathbf{Y}_{t-1} + \mathbf{1}_p)$. In summary, we have the following result; its proof is omitted since it uses identical arguments as those in the proof of Theorem 4.1. Note however that one of the main ingredients of the proof is to show that the score function (17) is a square integrable martingale; this fact is guaranteed by the conclusions of Lemma 3.2; in particular the fourth result.

Theorem 4.2. Consider model (4). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. Suppose that $\boldsymbol{\Theta}$ is compact and assume that the true value $\boldsymbol{\theta}_0$ belongs to the interior of $\boldsymbol{\Theta}$. Suppose that at the true value $\boldsymbol{\theta}_0$, the conditions of Proposition 3.3 hold true. Then there exists a fixed open neighborhood, say $O(\boldsymbol{\theta}_0)$, of $\boldsymbol{\theta}_0$ such that with probability tending to 1 as $n \to \infty$, the equation $S_n(\boldsymbol{\theta}) = 0$, where $S_n(\cdot)$ is defined by (17), has a unique solution, say $\widehat{\boldsymbol{\theta}}$. Furthermore, $\widehat{\boldsymbol{\theta}}$ is strongly consistent and asymptotically normal, as in Theorem 4.1, where the matrices $\mathbf{G}(\boldsymbol{\theta})$ and $\mathbf{H}(\boldsymbol{\theta})$ are defined by

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

and expectation is taken with respect to the stationary distribution of $\{\mathbf{Y}_t\}$.

Although the product form of (10) indicates independence, the dependence structure in (3) and (4) will be picked up explicitly through the dependence of (10) on the matrices $\mathbf{A}$ and $\mathbf{B}$. The copula structure, however, does not explicitly appear in (10), even though indirectly it does because of the conditional innovation $\mathbf{Y}_t \mid \boldsymbol{\lambda}_t$. (One could, of course, have chosen a more specific dependence model for these quantities. The copula was chosen because of its general way of describing dependence.) To recover the copula dependence one has to look at the conditional distribution of $\mathbf{Y}_t \mid \boldsymbol{\lambda}_t$ and compare it with the conditional distribution of $\mathbf{Y}_t^* \mid \boldsymbol{\lambda}_t$, say, generated by a suitable copula model conditional on $\boldsymbol{\lambda}_t$. There are several ways of comparing such distributions, e.g. the Kullback-Leibler or Hellinger distances. A thorough study of this problem requires a separate publication. In the supplement, we have opted for a preliminary and heuristic approach based on the newly developed concept of local Gaussian correlation; for more, including some simulation and real data evidence, see Sec. **??–??** in the supplement.

5. Simulation and data analysis

In this section we illustrate the theory by presenting a limited simulation study for the linear model. In addition we include a real data example. Further supporting material is given in the supplement in Sec **??**.

5.1. Simulations for the multivariate linear model

For the simulation study we only consider a two-dimensional process, that is $p = 2$. To initiate the maximization algorithm, we obtain starting values for the parameter vector $\boldsymbol{\theta} = (\mathbf{d}, \mathrm{vec}^T(\mathbf{A}), \mathrm{vec}^T(\mathbf{B}))$ as follows. We first fit a univariate model to each series by using the methods of [21] and [20]. Then, employing the univariate predictions obtained from each of the hidden process, we run a multivariate linear regression model by regressing the response to its lagged value and the vector of estimated hidden process. This method seems to work well in practice but further experiments are needed. Throughout the simulations we generate 1000 realizations with sample sizes of 500 and 1000 by employing the Clayton copula. We report the estimates of the parameters by averaging out the results from all simulations, and similarly, the standard errors correspond to the sampling standard errors of the estimates obtained by the simulation. Table 1 illustrates simulation results obtained from the linear model where the off-diagonal elements of the matrices $\mathbf{A}$ and $\mathbf{B}$ are non-zero, i.e. following parameters

$$\mathbf{A} = \begin{pmatrix} 0.3 & 0.05 \\ 0.1 & 0.25 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0.5 & 0.05 \\ 0.1 & 0.4 \end{pmatrix} \text{ and } \mathbf{d} = (0.5, 1). \tag{18}$$

Note that these parameter values yield $|||\mathbf{A} + \mathbf{B}|||_2 = 0.89 < 1$ but $|||\mathbf{A}|||_1 + |||\mathbf{B}|||_1 = 1$ (compare Propositions 3.1 and 3.2). The empirical results largely agree with the theoretical properties of the estimators for both values of the copula parameter ϕ with the exception of $\hat{\mathbf{d}}$ which does not approach normality satisfactorily, but the approximation improves for larger sample sizes. Further simulation results are given in the supplement.

Sample size	ϕ	$\hat{d}_1$	$\hat{d}_2$	$\hat{a}_{11}$	$\hat{a}_{22}$	$\hat{b}_{11}$	$\hat{b}_{22}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{b}_{12}$	$\hat{b}_{21}$
500	0	0.871	1.421	0.289	0.222	0.493	0.396	0.087	0.167	0.051	0.098
		(0.205)	(0.349)	(0.071)	(0.084)	(0.049)	(0.050)	(0.082)	(0.077)	(0.045)	(0.049)
	0.5	0.772	1.116	0.279	0.200	0.494	0.395	0.083	0.161	0.051	0.099
		(0.170)	(0.264)	(0.074)	(0.087)	(0.051)	(0.051)	(0.085)	(0.081)	(0.050)	(0.052)
1000	0	0.803	1.316	0.295	0.222	0.498	0.400	0.083	0.166	0.052	0.099
		(0.134)	(0.236)	(0.052)	(0.057)	(0.036)	(0.032)	(0.054)	(0.054)	(0.030)	(0.036)
	0.5	0.733	1.056	0.286	0.207	0.497	0.396	0.082	0.157	0.048	0.100
		(0.118)	(0.181)	(0.055)	(0.061)	(0.037)	(0.037)	(0.057)	(0.054)	(0.035)	(0.037)

Table 1. Simulation results for the multivariate linear model (1) by employing the Clayton copula with parameter ϕ. True parameter values are given by (18). Standard errors of the estimators are given in parentheses. Results are based on 1000 runs.

5.2. Real data analysis

As an illustration of this methodology, we fit the linear and log-linear models to a bivariate count time series which consists of the number of transactions per 15 seconds for the stocks Coca-Cola Company (KO) and IBM on September 19th 2005. The data are from the NYSE Trade and Quote (TAQ) database, that contains intraday transactions data for all securities listed on the New York Stock Exchange (NYSE). It is of interest to study how two heavily traded stocks in different sectors, influence each others trading activity. There are 1440 observations in each of the two series, covering trades from 09:30 to 16:30, excluding the first 15 minutes and last 15 minutes of transactions. We remove these data, because transaction counts (and all other measures of intraday activity such as, e.g., volume) are typically characterized by a U-shaped diurnal seasonality (more transactions at the open and close and less at midday), which can interfere with the measurement of auto- and cross-correlations, see, e.g. [32].

Figure 1 shows a time series plot of the data and Figure 2 depicts the autocorrelation function and cross-autocorrelation functions. Clearly, the plot of the autocorrelation functions reveals high correlation within and between the individual transaction series. Note further that mean number of transactions is 4.854 and 4.276, for IBM and KO stocks, respectively. The sample variances are 13.809 (IBM) and 10.707 (KO), that is the data clearly shows marginal overdispersion.

Table 2 shows estimated parameters after fitting the linear and log-linear models to these data. In both cases, the standard errors given in parentheses under the estimated parameters in Table 2 were computed using the robust estimator of the covariance matrix given by $\mathbf{H}_n(\hat{\theta})^{-1}\mathbf{G}_n(\hat{\theta})\mathbf{H}_n(\hat{\theta})^{-1}$ where $\mathbf{H}_n$ and $\mathbf{G}_n$ are given in equation (13) and (14), respectively. The magnitude of the standard errors shows that the feedback process should be considered in both models.

Fitted model	$\hat{d}_1$	$\hat{d}_2$	$\hat{a}_{11}$	$\hat{a}_{22}$	$\hat{b}_{11}$	$\hat{b}_{22}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{b}_{12}$	$\hat{b}_{21}$
Linear	0.388	0.348	0.625	0.611	0.126	0.145	0.015	0.103	0.062	0.035
	(1.110)	(0.713)	(0.173)	(0.001)	(0.001)	(0.148)	(0.005)	(0.001)	(0.004)	(0.005)
Log-linear	0.110	0.149	0.830	0.720	0.104	0.141	-0.008	-0.032	0.035	0.026
	(0.001)	(0.152)	(0.085)	(0.035)	(0.143)	(0.056)	(0.003)	(0.001)	(0.012)	(0.0005)

Table 2. Fit of the linear and log-linear model. Standard errors given in parentheses.

The predictions from both models are denoted by $\hat{Y}_{i,t} = \lambda_{i,t}(\hat{\theta})$ for $i = 1$ and 2, and are shown in Figure 1. We see that the predictions approximate the observed processes reasonably well. We compare the two models by calculating the RMSE using the predictions $\hat{Y}_{i,t}$ for $i = 1$ and 2 for both models. This gives an RMSE of 190.06 for the linear model and 193.25 for the log-linear model, indicating in total a better fit using the linear model. To examine the model fit, we consider the Pearson residuals, defined by $e_{i,t} = (Y_{i,t} - \lambda_{i,t})/\sqrt{\lambda_{i,t}}$ for $i = 1, 2$. Under the correct model, the sequence $e_{i,t}$ is a white noise sequence with constant variance. We substitute $\lambda_{i,t}$ by $\lambda_{i,t}(\hat{\theta})$ to obtain $\hat{e}_{i,t}$. We compute the Pearson residuals for

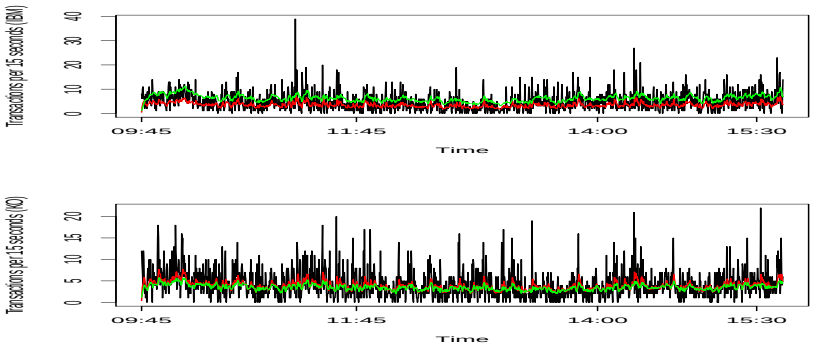

Figure 1: Number of transactions per 15 seconds for IBM (top) and Coca-Cola (bottom) and the respective predicted number of transactions from the linear model (red lines) and log-linear model (green lines).

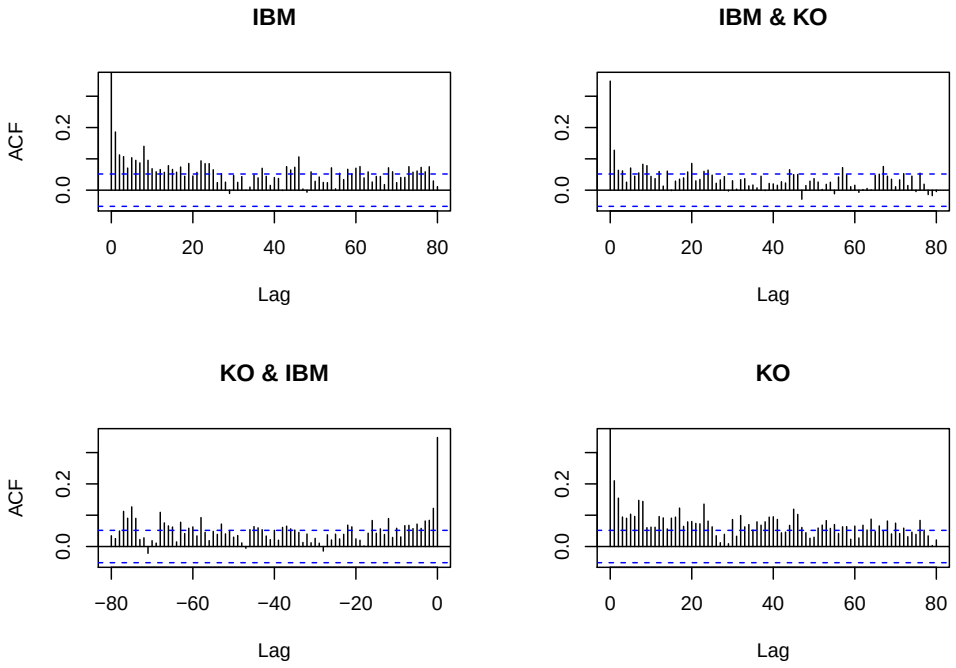

Figure 2: Auto- and cross-correlation function of the transaction data.

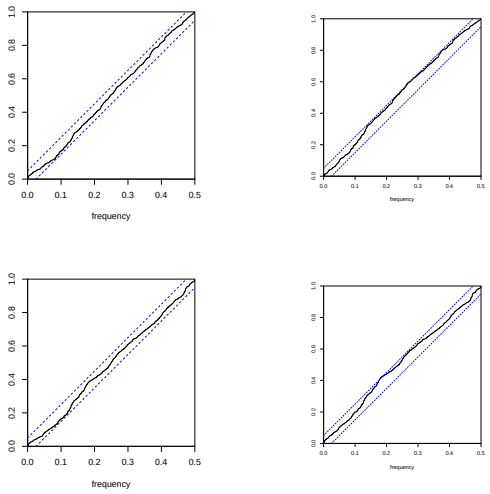

Figure 3: Left: Cumulative periodogram plots of the Pearson residuals from the linear fit of IBM (top) and Coca-Cola (bottom). Right: Cumulative periodogram plots of the Pearson residuals from the log-linear fit of IBM (top) and Coca-Cola (bottom).

both models, and examine their cumulative periodograms. Figure 3 supports the marginal whiteness of the residual process. A log-linear model that includes $\log(\mathbf{Y}_{t-1} + c\mathbf{1}_p)$ for some constant $c > 1$ could had been entertained for modelling these data. However, predictions obtained after fitting such model for various values of c did not alter our results considerably (see also [23, p.571] for the univariate case). Finally, the results of the copula estimation, for this data example, are reported in the supplement.

Acknowledgements

The authors would like to thank the Editor, Associate Editor and two reviewers for valuable comments and suggestions. Part of this research was carried out while K. Fokianos was at the Department of Mathematics & Statistics, University of Cyprus. This work was supported by the Institute of Advanced Studies of the University of Cergy Pontoise under the Paris Seine Initiative for Excellence ("Investissements d'Avenir" ANR-16-IDEX-0008) . In addition, it has been developed within the MME-DII center of excellence (ANR-11-LABEX-0023-01) and with the help of PAI-CONICYT MEC Nr 80170072. B. Støve thanks the Financial Market Fund (Norway) for support.

Appendix

It is easy to see that $\boldsymbol{\lambda}^\star = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{d}$ is a fixed point of the skeleton (3). The proof of the following lemma is quite analogous to the proof of [21, Lemma A.1] and it is omitted.

Lemma A-1. Let $\{\boldsymbol{\lambda}_t\}$ be a Markov chain defined by (4) or (5). If $|||\mathbf{A}|||_2 < 1$, then every point in $[\lambda_1^\star, \infty) \times \cdots \times [\lambda_p^\star, \infty)$ is reachable, where $\lambda_i^\star$ denotes the i'th component of the vector $\boldsymbol{\lambda}^\star$.

A-1. Proof of Proposition 3.1

The conditions of ϕ-irreducibility and the existence of small sets can be proved along the lines of the proof of [21, Prop. 2.1] provided that $|||\mathbf{A}|||_2 < 1$. As in the proof of that Proposition we use the Tweedie criterion to prove geometric ergodicity. Define now the test function $V(\mathbf{x}) = 1 + \|\mathbf{x}\|_2^r$. Then, we obtain as $\lambda_i \to \infty$, $i = 1, 2, \ldots, p$,

$$\begin{aligned} \mathrm{E}\left[V(\boldsymbol{\lambda}_t^m)|\boldsymbol{\lambda}_{t-1}^m = \boldsymbol{\lambda}\right] &= 1 + \mathrm{E}\left[\|\mathbf{d} + \mathbf{A}\boldsymbol{\lambda} + \mathbf{B}\mathbf{Y}_{t-1}^m + \boldsymbol{\epsilon}_{t;m}\|_2^r\right] \\ &\sim \mathrm{E}\left[\|\mathbf{A}\boldsymbol{\lambda} + \mathbf{B}\mathbf{Y}_{t-1}^m\|_2^2\right]^\mu, \end{aligned}$$

where we assume, without loss of generality, that $\mu = r/2$, r a positive integer. Next,

$$\mathrm{E}\left[\|\mathbf{A}\boldsymbol{\lambda} + \mathbf{B}\mathbf{Y}_{t-1}^m\|_2^r\right] = \mathrm{E}\left[\left[\sum_{i=1}^p \left((\mathbf{A}\boldsymbol{\lambda})_i + (\mathbf{B}\mathbf{Y}_{t-1}^m)_i\right)^2\right]\right]^\mu := \mathrm{E}\left(\sum_{i=1}^p C_i\right)^\mu,$$

where $(\mathbf{A}\boldsymbol{\lambda})_i$ and $(\mathbf{B}\mathbf{Y}_{t-1}^m)_i$ are the ith components of the vectors $\mathbf{A}\boldsymbol{\lambda}$ and $\mathbf{B}\mathbf{Y}_{t-1}^m$, respectively. But

$$\left(\sum_{i=1}^p C_i\right)^\mu = \sum_{i_1} \cdots \sum_{i_p} \frac{\mu!}{i_1! \ldots i_p!} C_1^{i_1} \ldots C_p^{i_p},$$

where the sum extends over all indices $i_j, j = 1, 2, \ldots p$ such that $\sum_{j=1}^p i_j = \mu$. Successive use of the Cauchy-Schwartz inequality yields

$$\mathrm{E}\left(C_1^{i_1} \ldots C_p^{i_p}\right) \le \mathrm{E}^{1/2l_1}(C_1^{2i_1 l_1}) \ldots \mathrm{E}^{1/2l_p}(C_p^{2i_p l_p}),$$

where $1 \le l_p \le 2^{p-2}$, and

$$\mathrm{E}\left(C_k^{2i_k l_k}\right) = \mathrm{E}\left[(\mathbf{A}\boldsymbol{\lambda})_k + (\mathbf{B}\mathbf{Y}_{t-1}^m)_k\right]^{4i_k l_k} = \mathrm{E}\left[\sum_{j=0}^{4i_k l_k} \binom{4i_k l_k}{j} (\mathbf{A}\boldsymbol{\lambda})_k^j (\mathbf{B}\mathbf{Y}_{t-1})_k^{4i_k l_k - j}\right].$$

But using the reasoning on page 26 of [22], as $\lambda_k \to \infty$, $k = 1, \ldots, p$,

$$\mathrm{E}\left[(\mathbf{B}\mathbf{Y}_{t-1})_k^{4i_k l_k - j} | \boldsymbol{\lambda}_{t-1} = \boldsymbol{\lambda}\right] \sim (\mathbf{B}\boldsymbol{\lambda})_k^{4i_k l_k - j}.$$

Hence $\mathrm{E}^{1/2l_k}\left(C_k^{2i_k l_k}\right) \sim ((\mathbf{A}+\mathbf{B})\boldsymbol{\lambda})_k^{2i_k}$, and asymptotically $\mathrm{E}\left(C_1^{i_1}\dots C_p^{i_p}\right) \leq ((\mathbf{A}+\mathbf{B})\boldsymbol{\lambda})_1^{2i_1}\dots((\mathbf{A}+\mathbf{B})\boldsymbol{\lambda})_p^{2i_p}$. Therefore we obtain that

$$\begin{aligned}\mathrm{E}\left(\sum_{i=1}^{p} C_i\right)^{\mu} &\leq \sum_{i_1}\dots\sum_{i_p}\frac{\mu!}{i_1!\dots i_p!}\left[((\mathbf{A}+\mathbf{B})\boldsymbol{\lambda})_1^2\right]^{i_1}\dots\left[((\mathbf{A}+\mathbf{B})\boldsymbol{\lambda})_p^2\right]^{i_p}\\ &= \left[\sum_{j=1}^{p}((\mathbf{A}+\mathbf{B})\boldsymbol{\lambda})_j^2\right]^{\mu} = \left(\|\|(\mathbf{A}+\mathbf{B})\boldsymbol{\lambda}\|\|_2^2\right)^{\mu} \leq \left(\|\|(\mathbf{A}+\mathbf{B})\|\|_2^2\|\boldsymbol{\lambda}\|_2^2\right)^{\mu}\end{aligned}$$

which, using the Tweedie criterion as in [21, Prop. 2.1], implies that $\|\|\mathbf{A}+\mathbf{B}\|\|_2 < 1$ is a sufficient condition, and the proposition thus holds.

A-2. Proof of Lemma 3.1

To prove the first item of the Lemma, note that

$$\begin{aligned}\|\mathrm{E}(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t)\|_2 &= \|\mathbf{A}\mathrm{E}(\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1}) + \mathbf{B}\mathrm{E}(\mathbf{Y}_{t-1}^m - \mathbf{Y}_{t-1}) + \mathrm{E}(\boldsymbol{\epsilon}_t^m)\|_2\\ &= \|\mathbf{A}\mathrm{E}\left(\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1}\right) + \mathbf{B}\left[\mathrm{E}\left[\mathrm{E}\left((\mathbf{Y}_{t-1}^m|\mathscr{F}_{t-1;m}^{\mathbf{Y},\boldsymbol{\lambda}})\right)\right] - \mathrm{E}\left[\mathrm{E}\left(\mathbf{Y}_{t-1}|\mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}\right)\right]\right] + \mathrm{E}(\boldsymbol{\epsilon}_t^m)\|_2\\ &\leq \|\|\mathbf{A}+\mathbf{B}\|\|_2\|\mathrm{E}(\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1})\|_2 + \|\mathrm{E}(\boldsymbol{\epsilon}_t^m)\|_2,\end{aligned}$$

where $\mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}$ and $\mathscr{F}_{t-1;m}^{\mathbf{Y},\boldsymbol{\lambda}}$ are the σ-algebras generated by $\{\boldsymbol{\lambda}_s, s \leq t\}$ and $\{\boldsymbol{\lambda}_s^m, s \leq t\}$, respectively. By recursion and the fact that $\|\mathrm{E}(\boldsymbol{\epsilon}_t^m)\|_2 \leq c_m$ which tends to zero as $m \to \infty$ we obtain the desired result. To prove the second statement, note that as $m \to \infty$,

$$\mathrm{E}\|(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t)\|_2^2 \sim \mathrm{E}\|\mathbf{A}(\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1}) + \mathbf{B}(\mathbf{Y}_{t-1}^m - \mathbf{Y}_{t-1})\|_2^2.$$

Let $\Delta_{t-1}\boldsymbol{\lambda} = \boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1}$ and $\Delta_{t-1}\mathbf{Y} = \mathbf{Y}_{t-1}^m - \mathbf{Y}_{t-1}$, then

$$\begin{aligned}\mathrm{E}\|(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t)\|_2^2 &\sim \mathrm{E}\left[\Delta_{t-1}\boldsymbol{\lambda}^T\mathbf{A}^T\mathbf{A}\Delta_{t-1}\boldsymbol{\lambda} + \Delta_{t-1}\boldsymbol{\lambda}^T\mathbf{A}^T\mathbf{B}\Delta_{t-1}\mathbf{Y} + \Delta_{t-1}\mathbf{Y}^T\mathbf{B}^T\mathbf{A}\Delta_{t-1}\boldsymbol{\lambda} + \Delta_{t-1}\mathbf{Y}^T\mathbf{B}^T\mathbf{B}\Delta_{t-1}\mathbf{Y}\right]\\ &= \mathrm{E}\left[\Delta_{t-1}\boldsymbol{\lambda}^T\mathbf{C}\Delta_{t-1}\boldsymbol{\lambda} + \Delta_{t-1}\boldsymbol{\lambda}^T\mathbf{D}\Delta_{t-1}\mathbf{Y} + \Delta_{t-1}\mathbf{Y}^T\mathbf{D}^T\Delta_{t-1}\boldsymbol{\lambda} + \Delta_{t-1}\mathbf{Y}^T\mathbf{E}\Delta_{t-1}\mathbf{Y}\right]\\ &:= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left[c_{ij}\Delta_{t-1}\lambda_i\Delta_{t-1}\lambda_j + d_{ij}\Delta_{t-1}\lambda_i\Delta_{t-1}Y_j + d_{ji}\Delta_{t-1}\lambda_i\Delta_{t-1}Y_j + e_{ij}\Delta_{t-1}Y_i\Delta_{t-1}Y_j\right],\end{aligned}$$

where $\mathbf{C} = \mathbf{A}^T\mathbf{A}$, $\mathbf{D} = \mathbf{A}^T\mathbf{B}$ and $\mathbf{E} = \mathbf{B}^T\mathbf{B}$. By using properties of conditional expectation as before, we obtain

$$\mathrm{E}\left[d_{ij}\Delta_{t-1}\lambda_i\Delta_{t-1}Y_j + d_{ji}\Delta_{t-1}\lambda_i\Delta_{t-1}Y_j\right] = \mathrm{E}\left[d_{ij}\Delta_{t-1}\lambda_i\Delta_{t-1}\lambda_j + d_{ji}\Delta_{t-1}\lambda_i\Delta_{t-1}\lambda_j\right]$$

In addition, following the proof in [21, Lemma 2.1], and using the above conditioning argument, $\mathrm{E}\left(\Delta_{t-1}Y_i^2\right) = \mathrm{E}\left(\Delta_{t-1}\lambda_i\right)^2 + 2\delta_{i,m}$, where $\delta_{i,m} \to 0$, as $m \to \infty$. For the cross-terms we have to condition on the copula structure, $\mathscr{F}_{t-1;m}^{\mathbf{Y},\boldsymbol{\lambda}}$, as well i.e.

$$\mathrm{E}\left(\Delta_{t-1}Y_i\Delta_{t-1}Y_j\right) \quad = \quad \mathrm{E}\left[\mathrm{E}\left[\Delta_{t-1}Y_i\Delta_{t-1}Y_j|\mathscr{F}_{t-1;m}^{\mathbf{Y},\boldsymbol{\lambda}},\mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}\right]\right] = \mathrm{E}\left(\Delta_{t-1}\lambda_i\Delta_{t-1}\lambda_j\right).$$

Collecting all previous results, we obtain

$$\mathrm{E}\|\left(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t\right)\|_2^2 = \mathrm{E}\|(\mathbf{A}+\mathbf{B})(\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1})\|_2^2 + D_m \leq |||\mathbf{A}+\mathbf{B}|||_2^2\mathrm{E}\|(\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1})\|_2^2 + D_m,$$

where $D_m \to 0$ as $m \to \infty$. The last two statements are proved using straightforward adaptation of the proof of [21, Lemma 2.1].

A-3. Proof of Proposition 3.2

The proof is based on [17, Thm. 3.1] and parallels the proof given by [15]. In proving weak dependence, we define the $\mathbf{X}_t = (\mathbf{Y}_t^T, \boldsymbol{\lambda}_t^T)^T$ and we employ the norm $\|\mathbf{x}\|_\epsilon = \|\mathbf{y}\|_1 + \epsilon\|\boldsymbol{\lambda}\|_1$, where ϵ is not necessarily small. Then, the contraction property is verified by noting that $\mathbf{X}_t = F(\mathbf{X}_{t-1}^T, \mathbf{N}_t^T)$ where $\mathbf{N}_t$ is an iid sequence of p-variate copula Poisson processes and choosing $\epsilon = |||\mathbf{A}|||_1/|||\mathbf{B}|||_1$. This proves that $\mathrm{E}[\|\mathbf{Y}_t\|_1] < \infty$ and $\mathrm{E}[\|\boldsymbol{\lambda}_t\|_1] < \infty$.

To show finiteness of moments we will be using induction and a different technique than the method used in [15]. More precisely, suppose that $\mathrm{E}[\|\mathbf{Y}_t\|_{r-1}^{r-1}] < \infty$ and $\mathrm{E}[\|\boldsymbol{\lambda}_t\|_{r-1}^{r-1}] < \infty$ for $r \in \mathbb{N}$ and $r > 1$. Then consider the i-th component of $\mathbf{Y}_t$. But

$$\begin{aligned}\mathrm{E}\left[Y_{i,t}^r \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}\right] &\leq \mathrm{E}\left[(Y_{i,t})_r \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}\right] + \sum_{k=1}^{r-1}|\delta_{ik}(r)|\mathrm{E}\left[Y_{i,t}^k \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}\right] \\ &= \lambda_{i,t}^r + \sum_{k=1}^{r-1}|\delta_{ik}(r)|\mathrm{E}\left[Y_{i,t}^k \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}\right],\end{aligned}$$

where $(x)_r = x(x-1)....(x-r+1)$, $\{\delta_{jk}(r),\ k = 1, 2, \ldots, r-1\}$ are some constants and the first line follows from properties of $(x)_r$ while the second line follows form properties of the Poisson distribution. By taking expectations and using the c_r-inequality, we obtain that

$$\mathrm{E}^{1/r}\left[Y_{i,t}^r\right] \leq \mathrm{E}^{1/r}\left[\lambda_{i,t}^r\right] + \sum_{k=1}^{r-1}|\delta_{ik}(r)|^{1/r}\mu_i,$$

where $\mu_i = \max_{k<r}\mathrm{E}\left[Y_{i,t}^k \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}}\right]$, which exists by the induction hypothesis. But

$$\mathrm{E}(\lambda_{i,t}^r) \leq \mathrm{E}(Y_{i,t}^r),$$

because of the properties of the linear model. Therefore, we obtain that (because of (3))

$$\mathrm{E}^{1/r}[Y_{i,t}^r] \leq d_i + \sum_{j=1}^{p} a_{ij}\mathrm{E}^{1/r}[Y_{i,t}^r] + \sum_{j=1}^{p} b_{ij}\mathrm{E}^{1/r}[Y_{i,t}^r] + \sum_{k=1}^{r-1} |\delta_{ik}(r)|^{1/r}\mu_i,$$

and by summing up, using the definition of $|||.|||_1$ and its properties, we obtain that

$$\sum_{i=1}^{p} \mathrm{E}^{1/r}[Y_{i,t}^r] \leq \sum_{i=1}^{p} d_i + (|||A|||_1 + |||B|||_1)\sum_{i=1}^{p} \mathrm{E}^{1/r}[Y_{i,t}^r] + \sum_{i=1}^{p}\sum_{k=1}^{r-1} |\delta_{ik}(r)|^{1/r}\mu_i.$$

A-4. Proof of Lemma 3.2

We will prove the second and fourth conclusion as the other results follow from [23] and the proof of Lemma 3.1. But to prove the second statement, note that

$$\begin{aligned}
\mathrm{E}\|(\boldsymbol{\nu}_t^m - \boldsymbol{\nu}_t)\|_2^2 &= \mathrm{E}\|\mathbf{A}\mathrm{E}(\boldsymbol{\nu}_{t-1}^m - \boldsymbol{\nu}_{t-1}) + \mathbf{B}\mathrm{E}(\log(\mathbf{Y}_{t-1}^m + \mathbf{1}_p) - \log(\mathbf{Y}_{t-1} + \mathbf{1}_p)) + \mathrm{E}(\boldsymbol{\epsilon}_t^m)\|_2^2 \\
&\leq |||\mathbf{A}|||_2^2\mathrm{E}\|\boldsymbol{\nu}_{t-1}^m - \boldsymbol{\nu}_{t-1}\|_2^2 + |||\mathbf{B}|||_2^2\mathrm{E}\|\log(\mathbf{Y}_{t-1}^m + \mathbf{1}_p) - \log(\mathbf{Y}_{t-1} + \mathbf{1}_p)\|_2^2 \\
&+ 2|||\mathbf{A}|||_2|||\mathbf{B}|||_2\sqrt{\mathrm{E}\|\boldsymbol{\nu}_{t-1}^m - \boldsymbol{\nu}_{t-1}\|_2^2\mathrm{E}\|\log(\mathbf{Y}_{t-1}^m + \mathbf{1}_p) - \log(\mathbf{Y}_{t-1} + \mathbf{1}_p)|_2^2} + \kappa c_m^2,
\end{aligned}$$

where $\kappa > 0$. Consider now the $\mathrm{E}(\log(Y_{j,t-1}^m + 1) - \log(Y_{j,t-1}))^2$, $j = 1, 2, \ldots, p$. Then, following the proof of [23, Lemma 2.1] and assuming without loss of generality that $\lambda_{j,t-1}^m \geq \lambda_{j,t-1}$ we obtain that $((Y_{j,t-1}^m + 1)/(Y_{j,t-1} + 1) \geq 1$. Therefore by using Jensen's inequality (by employing the function $(\log x)^2$) we obtain that

$$\mathrm{E}\left[\log\frac{Y_{j,t-1}^m + 1}{Y_{j,t-1} + 1}\right]^2 \leq \left[\log \mathrm{E}\left(\frac{Y_{j,t-1}^m + 1}{Y_{j,t-1} + 1}\right)^2\right].$$

But according to [23, p. 576] the right hand side of the above inequality is bounded by $\mathrm{E}(\nu_{j,t-1}^m - \nu_{j;t-1})^2$ for $j = 1, 2, \ldots, p$. Hence, the conclusion of the Lemma follows again by the same arguments used in the proof of Lemma 3.1.

To prove the fourth result, we follow [23, pp. 576-577]. Consider the test function $V(\mathbf{x}) = \exp(r\|\mathbf{x}\|_2)$ for $r \in \mathbb{N}$. Set $b = r|||\mathbf{B}|||_2$. Then

$$\mathrm{E}[\exp(r\|\boldsymbol{\nu}_t^m\|_2) \mid \boldsymbol{\nu}_{t-1}^m = \boldsymbol{\nu}] \leq \exp(r(\|\mathbf{d}\|_2 + |||\mathbf{A}|||_2\|\boldsymbol{\nu}\|_2))\mathrm{E}[\exp(r|||\mathbf{B}|||_2\|\log(\mathbf{Y}_{t-1}^m + \mathbf{1}_p)\|_2) \mid \boldsymbol{\nu}_{t-1}^m = \boldsymbol{\nu}].$$

However

$$\begin{aligned}
\mathrm{E}\Big[\exp\Big(b\|\log(\mathbf{Y}_{t-1}^m + \mathbf{1}_p)\|_2\Big)|\boldsymbol{\nu}_{t-1}^m = \boldsymbol{\nu}\Big] &= \mathrm{E}\left\{\exp\Big[b\Big(\sum_{i=1}^{p}\log^2(Y_{i,t-1}^m + 1)\Big)^{1/2}\Big]|\boldsymbol{\nu}_{t-1}^m = \boldsymbol{\nu}\right\} \\
&= \mathrm{E}\left\{\exp\Big[b\Big(\sum_{i=1}^{p}\Big(\nu_i + \Big(\frac{\log(Y_{i,t-1}^m + 1)}{\exp(\nu_i)}\Big)\Big)^2\Big)^{1/2}\Big]|\boldsymbol{\nu}_{t-1}^m = \boldsymbol{\nu}\right\}.
\end{aligned}$$

But

$$\mathrm{Var}\Big[\frac{Y_t^m+1}{\exp(\nu_i)} \mid \boldsymbol{\nu}_{t-1}^m=\boldsymbol{\nu}\Big]=\exp(-\nu_i)\to 0, \tag{A-1}$$

provided that $\nu_i\to\infty$ for all $i=1,2,\ldots,p$. Therefore we have that

$$\mathrm{Var}\Big[\log\Big(\frac{Y_t^m+1}{\exp(\nu_i)}\Big)\mid \boldsymbol{\nu}_{t-1}^m=\boldsymbol{\nu}\Big]\to 0,$$

by the delta-method for moments and provided that $\nu_i\to\infty$ for all $i=1,2,\ldots,p$. Using now the multivariate delta-method and Cauchy-Schwartz inequality to the function $g(x_1,\ldots,x_p)=\exp(b(\sum_i^p(\nu_i+x_i)^2)^{1/2})$ (with some abuse of notation), we obtain that

$$\mathrm{Var}\left\{\exp\Big[b\Big(\sum_{i=1}^p\Big(\nu_i+\Big(\frac{\log(Y_{i,t-1}^m+1)}{\exp(\nu_i)}\Big)\Big)^2\Big)^{1/2}\Big]|\boldsymbol{\nu}_{t-1}^m=\boldsymbol{\nu}\right\}\to 0.$$

However

$$\mathrm{E}\Big[\frac{Y_t^m+1}{\exp(\nu_i)} \mid \boldsymbol{\nu}_{t-1}^m=\boldsymbol{\nu}\Big]\sim 1 \tag{A-2}$$

provided that $\nu_i\to\infty$ for all $i=1,2,\ldots,p$. Therefore, asymptotically, we obtain that

$$\mathrm{E}\Big[\exp\Big(b\|\log(\mathbf{Y}_{t-1}^m+\mathbf{1}_p)\|_2\Big)|\boldsymbol{\nu}_{t-1}^m=\boldsymbol{\nu}\Big]\sim\exp(b\|\boldsymbol{\nu}\|_2).$$

To complete the proof, we note that the above calculations show that

$$\mathrm{E}\Big[\exp(r\|\boldsymbol{\nu}_t^m\|_2)|\boldsymbol{\nu}_{t-1}^m=\boldsymbol{\nu}\Big]\le\exp(r(\||\mathbf{A}_2\||_2+\||\mathbf{B}_2\||_2-1)\|\boldsymbol{\nu}\|_2)\exp(r\|\boldsymbol{\nu}\|_2).$$

Therefore, the conclusion follows as in [23, pp. 576-577].

A-5. Proof of Proposition 3.4

For the log-linear model we prove weak dependence by the following method. Set $Y_{j,t}=N_{j,t}(\exp(\nu_{j,t}))$, $j=1,2,\ldots,p$. Then setting $Z_{j,t}=\log(1+Y_{j,t})$ we have for $\mathbf{X}_t=(\mathbf{Z}_t,\boldsymbol{\nu}_t)$ with $\mathbf{Z}_t=(Z_{j,t},j=1,2,\ldots,p)$ and $N_t=(N_{j,t},j=1,2,\ldots,d)$ that

$$\mathbf{X}_t=(\mathbf{Z}_t,\boldsymbol{\nu}_t)=F(\mathbf{X}_{t-1}^T,\mathbf{N}_t^T),$$

where $\mathbf{N}_t=(N_{j,t},j=1,2,\ldots,p)$ iid copula p-variate Poisson processes. Then using again the same arguments as in [15] we obtain (with the same norm) that

$$\begin{aligned}\mathrm{E}\left[\|F(\mathbf{x},\mathbf{N})-F(\mathbf{x}^\star,\mathbf{N})\|_\epsilon\right] &\le \sum_{j=1}^p\|\Big(\mathbf{A}(\boldsymbol{\nu}-\boldsymbol{\nu}^\star)\Big)_j\|_1+\sum_{j=1}^p\|\Big(\mathbf{B}(\boldsymbol{\zeta}-\boldsymbol{\zeta}^\star)\Big)_j\|_1\\ &+ \epsilon\Big(\|\mathbf{A}(\nu-\nu^\star)\|_1+\|\mathbf{B}(\boldsymbol{\zeta}-\boldsymbol{\zeta}^\star)\|_1\Big)\\ &\le (1+\epsilon)\Big(\||\mathbf{A}\||_1\|\boldsymbol{\nu}-\boldsymbol{\nu}^\star\|_1+\||\mathbf{B}\||_1\|\boldsymbol{\zeta}-\boldsymbol{\zeta}^\star\|_1\Big),\end{aligned}$$

where the first inequality follows from [23, pp.575–576]. The results now follow as in [15]. Now we show existence of moments for the log-linear model. Suppose that $r \in \mathbb{N}$. Then

$$\mathrm{E}[\exp(r\|\boldsymbol{\nu}_t\|_1) \mid \boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}] \quad \leq \quad \exp(r(\|\mathbf{d}\|_1 + |||\mathbf{A}|||_1\|\boldsymbol{\nu}\|_1))\mathrm{E}[\exp(r|||\mathbf{B}|||_1\|\mathbf{Z}_{t-1}\|_1) \mid \boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}]$$

With $b = r|||\mathbf{B}|||_1$, for the second factor of the right hand side we obtain that

$$\mathrm{E}\Big[\exp\Big(b\|\mathbf{Z}_{t-1}\|_1\Big) \mid \boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}\Big] = \exp(b\|\boldsymbol{\nu}\|_1)\mathrm{E}\Big[\prod_{i=1}^{p}\Big(\frac{Y_{i,t-1}+1}{\exp(\nu_i)}\Big)^b \mid \boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}\Big]$$

But from the proof of Lemma 3.2 (see eq. (A-1)) and using similar arguments

$$\mathrm{Var}\Big[\prod_{i=1}^{p}\Big(\frac{Y_{i,t-1}+1}{\exp(\nu_i)}\Big)^b \mid \boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}\Big] \to 0,$$

provided that $\nu_i \to \infty$, for all $i = 1, 2 \ldots, p$. In addition, because of (A-2) and the multivariate delta-method of moments

$$\mathrm{E}\Big[\prod_{i=1}^{p}\Big(\frac{Y_{i,t-1}+1}{\exp(\nu_i)}\Big)^b \mid \boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}\Big] \to 1,$$

provided that $\nu_i \to \infty$, for all $i = 1, 2 \ldots, p$. The above two displays show that

$$\mathrm{E}\Big[\exp\Big(b\|\mathbf{Z}_{t-1}\|_1\Big) \mid \boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}\Big] \sim \exp(b\|\boldsymbol{\nu}\|_1),$$

as required.

A-6. Proof of Lemma 4.1

In what follows we drop notation that depends on $\boldsymbol{\theta}$ because all quantities are evaluated at the true parameter $\boldsymbol{\theta}_0$. The notation C refers to a generic constant. Initially, we show that

$$\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \mathbf{d}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \mathbf{d}^T}\Big|\Big|\Big|_2 < \gamma_m, \ a.s, \tag{A-3}$$

for some positive sequence $\gamma_m \to 0$, as $m \to \infty$. Using the first equation of (12) we obtain that

$$\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \mathbf{d}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \mathbf{d}^T}\Big|\Big|\Big|_2 \quad \leq \quad |||\mathbf{A}|||_2\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_{t-1}^m}{\partial \mathbf{d}^T} - \frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \mathbf{d}^T}\Big|\Big|\Big|_2$$

and therefore, by repeated substitution, (A-3) follows since $|||\mathbf{A}|||_2 < 1$ and the results of Lemma 3.1. Similarly,

$$\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \,\mathrm{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \,\mathrm{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2 \leq \gamma_m, \ a.s. \tag{A-4}$$

Indeed, using the second equation of (12), we obtain that

$$\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2 \leq \sqrt{p}\|\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1}\|_2 + |||\mathbf{A}|||_2 \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_{t-1}^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2,$$

where the first bound comes from the fact that in terms of the Frobenius matrix norm $|||\mathbf{I}_p|||_F = \sqrt{p}$. Therefore, by Lemma 3.1 we obtain the desired result. Finally, it can be shown quite analogously (by using again Lemma (3.1)) that

$$\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{B})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{B})}\Big|\Big|\Big|_2 \leq \gamma_m, \ a.s. \tag{A-5}$$

To prove the lemma, we consider the $d \times d$ matrix difference

$$\begin{aligned} \Big|\Big|\Big|s_t^m (s_t^m)^T - s_t s_t^T\Big|\Big|\Big|_2 &= \Big|\Big|\Big|(s_t^m - s_t)(s_t^m)^T + s_t(s_t^m - s_t)^T\Big|\Big|\Big|_2 \\ &\leq \|s_t^m - s_t\|_2 \|(s_t^m)^T\|_2 + \|s_t\|_2 \|(s_t^m - s_t)^T\|_2. \end{aligned} \tag{A-6}$$

But

$$\begin{aligned} s_t^m - s_t &= \Big[\Big(\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \boldsymbol{\theta}^T}\Big)^T - \Big(\frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big)^T\Big]\Big(\mathbf{D}_t^m\Big)^{-1}\Big(\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\Big) \\ &+ \Big(\frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big)^T\Big[\Big(\mathbf{D}_t^m\Big)^{-1} - \Big(\mathbf{D}_t\Big)^{-1}\Big]\Big(\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\Big) \\ &+ \Big(\frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big)^T \mathbf{D}_t^{-1}\Big[\Big(\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\Big) - \Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t\Big)\Big] \\ &= (I) + (II) + (III), \end{aligned} \tag{A-7}$$

with obvious notation. Then we obtain for the first term (I) of (A-7)

$$\|\Big[\Big(\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \boldsymbol{\theta}^T}\Big)^T - \Big(\frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big)^T\Big]\Big(\mathbf{D}_t^m\Big)^{-1}\Big(\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\Big)\|_2 \leq \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \boldsymbol{\theta}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big|\Big|\Big|_2 \Big|\Big|\Big|\Big(\mathbf{D}_t^m\Big)^{-1}\Big|\Big|\Big|_2 \|\Big(\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\Big)\|_2. \tag{A-8}$$

We deal with the first factor. Recall that $|||.|||_F$ stands for the Frobenius norm of a matrix. Then

$$\begin{aligned} \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \boldsymbol{\theta}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big|\Big|\Big|_2^2 &\leq \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \boldsymbol{\theta}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big|\Big|\Big|_F^2 \\ &= \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \mathbf{d}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \mathbf{d}^T}\Big|\Big|\Big|_F^2 + \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{A})}^2\Big|\Big|\Big|_F + \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{B})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{B})}\Big|\Big|\Big|_F^2 \\ &\leq p\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \mathbf{d}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \mathbf{d}^T}\Big|\Big|\Big|_2^2 + p^2\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2^2 + p^2\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{B})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{B})}\Big|\Big|\Big|_2^2, \end{aligned}$$

where the first and third inequality hold because of result 4.67(a) of [51] and the second inequality is a consequence of the definition of Frobenius norm. Then we need to show that

$$\mathrm{E}\Big[\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \mathbf{d}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \mathbf{d}^T}\Big|\Big|\Big|_2^2\Big], \ \mathrm{E}\Big[\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2^2\Big], \ \mathrm{E}\Big[\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{B})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{B})}\Big|\Big|\Big|_2^2\Big] \leq \gamma_m, \tag{A-9}$$

with $\gamma_m \to 0$. We deal with the middle term only; similar arguments can be used for the other two terms. Squaring the expression after (A-4) and taking expectations we obtain that

$$\begin{aligned}
\mathrm{E}\Big[\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2^2\Big] &\leq p\mathrm{E}\Big[\|\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1}\|_2^2\Big] + |||\mathbf{A}|||_2^2 \mathrm{E}\Big[\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_{t-1}^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2^2\Big] \\
&+ 2\sqrt{p}|||\mathbf{A}|||_2 \mathrm{E}\Big[\|\boldsymbol{\lambda}_{t-1}^m - \boldsymbol{\lambda}_{t-1}\|_2 \Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_{t-1}^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2\Big] \\
&\leq \delta_{2;m} + |||\mathbf{A}|||_2^2 \mathrm{E}\Big[\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_{t-1}^m}{\partial \operatorname{vec}^T(\mathbf{A})} - \frac{\partial \boldsymbol{\lambda}_{t-1}}{\partial \operatorname{vec}^T(\mathbf{A})}\Big|\Big|\Big|_2^2\Big] \\
&+ 2C\sqrt{p}|||\mathbf{A}|||_2 \sqrt{\delta_{2;m}} \leq \gamma_m,
\end{aligned}$$

where γ_m can become arbitrarily small. This follows from Proposition 3.1, (A-4) and the fact that $|||\mathbf{A}|||_2 < 1$. For the second term of (A-8) we note that

$$\Big|\Big|\Big|\Big(\mathbf{D}_t^m\Big)^{-1}\Big|\Big|\Big|_2 \leq \sqrt{p \max_{1\leq i\leq p} \frac{1}{d_i^2}} \leq C, \tag{A-10}$$

where d_i is the i'th component of $\mathbf{d}$. In addition

$$\mathrm{E}\Big[\|\Big(\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\Big)\|_2^2\Big] = \sum_{i=1}^p \mathrm{E}[\lambda_{i,t}^m] < C \tag{A-11}$$

by Proposition 3.1 and using a conditioning argument. Collecting (A-9), (A-10) and (A-11) an application of Cauchy-Schwartz inequality shows that the

$$\mathrm{E}\Big[\Big|\Big|\Big|\frac{\partial \boldsymbol{\lambda}_t^m}{\partial \boldsymbol{\theta}^T} - \frac{\partial \boldsymbol{\lambda}_t}{\partial \boldsymbol{\theta}^T}\Big|\Big|\Big|_2 \Big|\Big|\Big|\Big(\mathbf{D}_t^m\Big)^{-1}\Big|\Big|\Big|_2 \|\Big(\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\Big)\|_2\Big] \to 0,$$

as $m \to \infty$. Now we look at the second summand (II) of (A-7). First of all, we note that $\mathrm{E}\Big|\Big|\Big|\partial \boldsymbol{\lambda}_t / \partial \boldsymbol{\theta}^T\Big|\Big|\Big|_2^4 < C$. This is proved by using the same decomposition of the norm as the sum of norms of the matrix of derivatives with respect to $\mathbf{d}$, $\operatorname{vec}(\mathbf{A})$ and $\operatorname{vec}(\mathbf{B})$. Then using (12), the fact that $\Big|\Big|\Big|\mathbf{A}\Big|\Big|\Big|_2 < 1$ and the compactness of the parameter space, the result follows. In addition, for some finite constants (c_{ij}), we obtain that

$$\begin{aligned}
\mathrm{E}\Big[\|\mathbf{Y}_t^m - \boldsymbol{\lambda}_t^m\|_2^4\Big] &= \mathrm{E}\Big[\Big(\sum_{i=1}^p (Y_{i,t}^m - \lambda_{i,t}^m)^2\Big)^2\Big] \\
&= \mathrm{E}\Big[\sum_{i=1}^p (Y_{i,t}^m - \lambda_{i,t}^m)^4 + 2\sum_{i\neq j} (Y_{i,t}^m - \lambda_{i,t}^m)^2 (Y_{j,t}^m - \lambda_{j,t}^m)^2\Big] \\
&\leq \sum_{i=1}^p \mathrm{E}[(\lambda_{i,t}^m)^4] + \sum_{i=1}^p \sum_{j=1}^3 c_{ij}\mathrm{E}[(\lambda_{i,t}^m)^j] < C,
\end{aligned}$$

because of Proposition 3.1 and from the same arguments given in the proof of Proposition 3.2. Now we have that

$$\left|\left|\left|\left(\mathbf{D}_t^m\right)^{-1}-\left(\mathbf{D}_t\right)^{-1}\right|\right|\right|_2^2 \leq C\|\boldsymbol{\lambda}_t^m-\boldsymbol{\lambda}_t\|_2^2$$

and therefore its expected value tends to zero by Lemma 3.1. Collecting all these results we have that the expected value of (II) in (A-7) tends to zero. Finally, the expected value of term (III) in (A-7) tends to zero, as $m\to\infty$ by combining the above results and using Cauchy-Schwartz inequality and Lemma 3.1.In addition, the above results show that $\mathrm{E}[\|s_t\|_2^2]<\infty$. The conclusion of the Lemma follows.

References

[1] A. Ahmad. *Contributions à l'économétrie des séries temporelles à valeurs entières*. PhD thesis, University Charles De Gaulle-Lille III, France, 2016.

[2] A. Ahmad and C. Franq. Poisson QMLE of count time series models. *Journal of Time Series Analysis*, 37:291–314, 2016.

[3] C. M. Andreassen. *Models and inference for correlated count data*. PhD thesis, Aaarhus University, Denmark, 2013.

[4] D. Andrews. Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21:930–934, 1984.

[5] I. V. Basawa and B. L. S. Prakasa Rao. *Statistical Inference for Stochastic Processes*. Academic Press, London, 1980.

[6] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

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

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

[9] R. A. Davis, W. T. M. Dunsmuir, and Y. Wang. On autocorrelation in a Poisson regression model. *Biometrika*, 87:491–505, 2000.

[10] R. A. Davis and H. Liu. Theory and inference for a class of observation-driven models with application to time series of counts. *Statistica Sinica*, 26:1673–1707, 2016.

[11] J. Dedecker, P. Doukhan, G. Lang, José R. León R., S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190 of *Lecture Notes in Statistics*. Springer, New York, 2007.

[12] M. Denuit and P. Lambert. Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93:40–57, 2005.

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

[14] R. Douc, K. Fokianos, and E. Moulines. Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electron. J. Stat.*, 11:2707–2740, 2017.

[15] P. Doukhan, K. Fokianos, and D. Tjøstheim. On weak dependence conditions for Poisson autoregressions. *Statistics & Probability Letters*, 82:942–948, 2012. with a correction in Vol. 83, pp. 1926-1927.

[16] P. Doukhan and S. Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84:313–342, 1999.

[17] P. Doukhan and O. Wintenberger. Weakly dependent chains with infinite memory. *Stochastic Processes and Their Applications*, 118:1997–2013, 2008.

[18] B. Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81:709–721, 1986.

[19] R. Ferland, A. Latour, and D. Oraichi. Integer–valued GARCH processes. *Journal of Time Series Analysis Analysis*, 27:923–942, 2006.

[20] K. Fokianos. Statistical Analysis of Count Time Series Models: A GLM perspective. In R. Davis, S. Holan, R. Lund, and N. Ravishanker, editors, *Handbook of Discrete-Valued Time Series*, Handbooks of Modern Statistical Methods, pages 3–28. Chapman & Hall, London, 2015.

[21] K. Fokianos, A. Rahbek, and D. Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104:1430–1439, 2009.

[22] K. Fokianos, A. Rahbek, and D. Tjøstheim. Poisson autoregression (complete vesrion). 2009. available at http://pubs.amstat.org/toc/jasa/104/488.

[23] K. Fokianos and D. Tjøstheim. Log–linear Poisson autoregression. *Journal of Multivariate Analysis*, 102:563–578, 2011.

[24] C. Francq and J.-M. Zakoïan. *GARCH models: Stracture, Statistical Inference and Financial Applications*. Wiley, United Kingdom, 2010.

[25] J. Franke and T. Subba Rao. Multivariate first-order integer values autoregressions. Technical report, Department of Mathematics, UMIST, 1995.

[26] C. Genest and J. Nešlehová. A primer on copulas for count data. *Astin Bull.*, 37:475–515, 2007.

[27] A. Heinen and E. Rengifo. Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance*, 14:564 – 583, 2007.

[28] C. C. Heyde. *Quasi-Likelihood and its Applications: A General Approach to Optimal Parameter Estimation*. Springer, New York, 1997.

[29] H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, London, 1997.

[30] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete multivariate distributions*. John Wiley, New York, 1997.

[31] B. Jørgensen, S. Lundbye-Christensen, P. X.-K. Song, and L. Sun. State-space models for multivariate longitudinal data of mixed types. *The Canadian Journal of Statistics*, 24:385–402, 1996.

[32] R. Jung, R. Liesenfeld, and J.-F. Richard. Dynamic factor models for multivariate count data: an application to stock–market trading activity. *Journal of Business & Economic Statistics*, 29:73–85, 2011.

[33] D. Karlis. Modelling multivariate times series for counts. In R. Davis, S. Holan, R. Lund, and N. Rav-

ishanker, editors, *Handbook of Discrete-Valued Time Series*, Handbooks of Modern Statistical Methods, pages 407–424. Chapman & Hall, London, 2016.

[34] L. A. Klimko and P. I. Nelson. On conditional least squares estimation for stochastic processes. *The Annals of Statistics*, 6:629–642, 1978.

[35] A. Latour. The multivariate GINAR(p) process. *Advances in Applied Probability*, 29:228–248, 1997.

[36] Y. Lee, S. Lee, and D. Tjøstheim. Asymptotic normality and parameter change test for bivariate Poisson INGARCH models. *TEST*, 27:52–69, 2018.

[37] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[38] H. Liu. *Some models for time series of counts*. PhD thesis, Columbia University, USA, 2012.

[39] J. Livsey, Lund. R, S. Kechagias, and V. Pipiras. Multivariate integere-valued time series with flexible autocovariances and their application to major hurricane counts. *Annals of Applied Statistics*, 12:408–431, 2018.

[40] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, 2005.

[41] P. McCullagh and J. A Nelder. *Generalized Linear Models*. Chapman & Hall, London, 2nd edition, 1989.

[42] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993.

[43] M. Neumann. Absolute regularity and ergodicity of Poisson count processes. *Bernoulli*, 17:1268–1284, 2011.

[44] A. K. Nikoloulopoulos. On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference*, 143:1923–1937, 2013.

[45] M. Paul, L. Held, and A. M. Toschke. Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27:6250–6267, 2008.

[46] X. Pedeli and D. Karlis. On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis*, 34:206–220, 2013.

[47] X. Pedeli and D. Karlis. Some properties of multivariate INAR(1) processes. *Computational Statistics & Data Analysis*, 67:213 – 225, 2013.

[48] N. Ravishanker, V. Serhiyenko, and M. R. Willig. Hierarchical dynamic models for multivariate times series of counts. *Statistics and its Interface*, 7:559–570, 2014.

[49] N. Ravishanker, R. Venkatesan, and S. Hu. Dynamic models for time series of counts with a marketing application. In R. Davis, S. Holan, R. Lund, and N. Ravishanker, editors, *Handbook of Discrete-Valued Time Series*, Handbooks of Modern Statistical Methods, pages 425–446. Chapman & Hall, London, 2015.

[50] T. H. Rydberg and N. Shephard. A modeling framework for the prices and times of trades on the New York stock exchange. In W. J. Fitzgerlad, R. L. Smith, A. T. Walden, and P. C. Young, editors, *Nonlinear and Nonstationary Signal Processing*, pages 217–246. Isaac Newton Institute and Cambridge University Press, Cambridge, 2000.

[51] G. A. F. Seber. *A Matrix Handbook for Statisticians*. Wiley, Hoboken, NJ, 2008.

[52] M. Taniguchi and Y. Kakizawa. *Asymptotic theory of statistical inference for time series*. Springer, New

York, 2000.

[53] D. Tjøstheim. Some recent theory for autoregressive count time series. *TEST*, 21:413–438, 2012.

[54] D. Tjøstheim. Count Time Series with Observation-Driven Autoregressive Parameter Dynamics. In R. Davis, S. Holan, R. Lund, and N. Ravishanker, editors, *Handbook of Discrete-Valued Time Series*, Handbooks of Modern Statistical Methods, pages 77–100. Chapman & Hall, London, 2015.

[55] C. Wang, H. Liu, J.-F. Yao, R. A. Davis, and W. K. Li. Self-excited threshold Poisson autoregression. *Journal of the American Statistical Association*, 109:777–787, 2014.

[56] D. W. Woodard, D. S. Matteson, and S. G. Henderson. Stationarity of count-valued and nonlinear time series models. *Electronic Journal of Statistics*, 5:800–828, 2011.

[57] W. B. Wu. Asymptotic theory for stationary processes. *Statistics and Its Interface*, 4:207–226, 2011.

Received May 2017 and revised May 2019