Mixtures of Nonlinear Poisson Autoregressions

Paul Doukhan[1,2] Konstantinos Fokianos [3] Joseph Rynkiewicz[4]

[1]AGM, UMR 8088, University of Cergy-Pontoise
[2] CIMFAV, University of Valparaiso
e-mail: doukhan@u-cergy.fr

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

[4]SAMM, Université de Paris 1
e-mail: joseph.rynkiewicz@univ-paris1.fr

First Version: July 2018

Abstract

We study nonlinear mixtures of integer-valued ARCH type models for count time series data. We investigate the theoretical properties of these processes and we prove ergodicity and stationarity, under minimal assumptions. The model can be generalized by including a GARCH component but we show that such inclusion can be accommodated by an ARCH model whose number of lagged variables tend to infinity. This work complements some previous studies in this area and improves on existing results by developing asymptotic properties of the maximum likelihood estimator. Furthermore, we discuss the estimation problem when the number of mixtures regimes is overestimated and we prove theoretically that proper likelihood penalization enables asymptotic estimation of the true number of mixture regimes. A real data example illustrates the methodology.

Keywords: Bayesian Information Criterion, ergodicity, identifiability, nonlinear time series, penalized likelihood, stationarity.

1 Introduction

We consider models for count time series, which allow for the mean process to change according to the values of an unobservable discrete random variable. Equivalently, an unobservable discrete random variable determines the so called hidden regimes where the process alternates. In each of these regimes we assume that the conditional distribution of the time series given its past is Poisson. This is a standard distributional assumption which has been used for count time series analysis by several authors including Rydberg and Shephard (2000), Ferland et al. (2006) and Fokianos et al. (2009), for instance. To motivate this research, consider the top graph of Figure 2 which shows weekly number of E.coli cases in the state of North Rhine-Westphalia from January 2001 to May 2013. The plot indicates that there might be two regimes where the process of infected cases alternates. Indeed, we observe larger variability towards the end of the series. In this paper, we show how to define properly models to accommodate these features of data. In addition we study their properties, develop likelihood theory and suggest penalized likelihood methodology for obtaining a consistent estimator of the number of regimes.

Count time series analysis have received considerable attention, see Kedem and Fokianos (2002, Sec 4 & 5) and the recent edited volume by Davis et al. (2016) for several references. In the context of generalized linear models (see McCullagh and Nelder (1989)) the Poisson distribution has been widely used for modeling and inference of count data. In particular, INARCH (INteger ARCH) and INGARCH (INteger GARCH) processes have been found quite useful in applications; see the previous references for more on applied work. However application of mixtures models for count time series has not received a lot of attention. In the case of continuous valued time series some early work on mixtures for time series can be found in Jacobs et al. (1991), Le et al. (1996) and Francq and Zaköian (2001), for instance. In addition, Wong and Li (2000, 2001) have used a two-component mixture model with logistic weights that may depend on time and exogenous variables to model AR and ARCH models, respectively. Parameter estimation was performed via an EM algorithm and BIC (Bayesian information criterion) was employed for autoregressive lag selection. Stability properties of such models have been studied by Saikkonen (2007). Mixtures of AR models, in the Bayesian framework, have been considered by the recent contribution of Wood et al. (2011), among others.

In the context of count time series that we consider, some early work can be found in Albert (1991). In addition, Carvalho and Tanner (2005, 2007) proposed mixture-of-experts approach to model nonlinearities in time series models following generalized linear models. These authors applied maximum likelihood estimation, investigated identifiability and asymptotic normality of the estimates, and used AIC (Akaike Information Criterion) and BIC for selecting the number of components in the model. However, the conditions that are imposed for asymptotic inference are quite strict and their study focus on a log-linear model for count time series. A linear model was considered by Zhu et al. (2010) but the authors again did not provide some theoretical evidence regarding its properties. More recently, Berentsen et al. (2018) applied a Markov-switching Poisson loglinear model autoregressive model to a study of corporate defaults.

The aim of this paper is to give theoretical justification to some of the previous findings. In doing so, we extend previous studies because we consider a general class of nonlinear models. Indeed, the main model we consider is a mixture of Poisson models but the mean is allowed to depend nonlinearly on lagged values of the observed time series. Section 2 introduce the Mixture INARCH and INGARCH models. We discuss in detail the properties of both of these models and we show that the MINGARCH model can be approximated by a MINARCH model of infinite order. This result is of independent interest and is quite useful for the analysis of count time series models. Section 3 studies the properties of maximum likelihood estimator for the MINARCH model of finite order. Here we consider two cases; the case of known regimes and the case of unknown regimes. In the former case, it is shown that the maximum likelihood estimator of the vector of unknown parameters is strongly consistent and asymptotically normally distributed. This is a direct consequence of the regularity of the model in the case of known number of regimes. Furthermore, we study the asymptotic behavior of the likelihood ratio test statistic when the number of regimes is overestimated. This study yields a general criterion for selecting the true number of regimes. It turns out that the BIC satisfies this criterion showing that it can provide a consistent estimator for the unknown number of regimes. Section 4 discusses the challenges associated with estimation for the MINGARCH model. Section 5 gives a data analysis example and the paper concludes with an Appendix that contains proofs of all results.

2 Poisson Models for Mixtures of Count Time Series

Denote by Y a random variable whose probability mass function (pmf) is given by

$$\mathrm{P}[Y=y]=\sum_{k=1}^{K} p_k \frac{\exp(-\lambda_k)\lambda_k^y}{y!}, \quad y=0,1,2,\ldots, \tag{1}$$

where $\lambda_k > 0$, $0 < p_k < 1$, for $k = 1,2,\ldots,K$ such that $\sum_{k=1}^{K} p_k = 1$. Then we say that the distribution of Y is a K-class mixture of Poisson distributions. For ease of notation, we will denote the pmf (1) as $\mathrm{MP}(\boldsymbol{p},\boldsymbol{\lambda})$ with $\boldsymbol{p}=(p_1,\ldots,p_K)^T$ and $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_K)^T$. Elementary calculations show that

$$\mathrm{E}[Y]=\sum_{k=1}^{K} p_k\lambda_k, \quad \mathrm{Var}[Y]=\sum_{k=1}^{K} p_k\lambda(1+\lambda_k)-\Big(\sum_{k=1}^{K} p_k\lambda_k\Big)^2. \tag{2}$$

The Poisson mixture distribution (1) will be employed for modeling and analysis of integer valued count time series.

2.1 The MINARCH model

To fix notation, suppose that $\{Y_t\}$ denotes a count time series and let $\{\lambda_{k,t}\}$ be sequence of Poisson mean rates for $k = 1,2,\ldots,K$. Define $\boldsymbol{\lambda}_t=(\lambda_{1,t},\ldots,\lambda_{K,t})^T$. Suppose that $\mathcal{F}_t$ denotes the σ-field generated by $\{Y_s, s \le t\}$.

Let L be a positive integer and consider the following general class of nonlinear mixture models defined by

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t), \qquad \lambda_{k,t} = f_k(Y_{t-1}, \ldots, Y_{t-L}), \;\; k = 1, \ldots, K, \tag{3}$$

where $f_k(\cdot)$ are functions known up to a finite dimensional parameter vector, they are defined on $\mathbb{N}^L$ and take values on $(0, \infty)$, for all $k = 1, \ldots, K$. In this setup, several examples of nonlinear count time series models have been studied by Fokianos et al. (2009), Neumann (2011), Fokianos and Tjøstheim (2012), Doukhan et al. (2012) and Christou and Fokianos (2014). Indeed, if $p_1 = 1$ and $p_k = 0$, for $k \neq 1$, then (3) reduces to a Poisson nonlinear count time series model. Because of (1), it is easy to see that (3) implies that the conditional pmf of Y_t, given its past values, is given by

$$\text{P}[Y_t = y \,|Y_{t-1}, \ldots Y_{t-L}] = \sum_{k=1}^{K} p_k \frac{\exp(-f_k(Y_{t-1}, \ldots Y_{t-L}))f_k(Y_{t-1}, \ldots Y_{t-L})^y}{y!}.$$

We call (3) a nonlinear Mixture INteger ARCH model of order L and we denote this process by MINARCH(L).

Example 2.1 The simplest example of model (3) is given by assuming that the functions $f_k(\cdot)$ are linear on Y_{t-1}. In this case, we obtain that

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t), \quad \lambda_{k,t} = \psi_{k,0} + \psi_{k,1} Y_{t-1}, \tag{4}$$

where the unknown parameters are such that $\psi_{0,k}, \psi_{1,k} > 0$ for $k = 1, 2, \ldots, K$. The left plot of Figure 1 shows five hundred realizations of model (4) with $K = 2$ regimes and the estimated sample autocorrelation function. The plots illustrate that correlation decreases fast and that the process is taking values in two regimes; the one regime corresponds to low count values and the other regime to higher values. This is a consequence of the parameter specification. Indeed, for $\psi_{1,1} = 1/4$ we expect this regime to be the one which corresponds to low values of count data. In contrast, $\psi_{2,1} = 6/5 > 1$ implies that in the second region the process will assume larger values. Note that we allow for the autoregressive coefficient that corresponds to the second regime to be larger than 1. This is a condition which does not guarantee stationarity for the case of a simple linear Poisson model, see Ferland et al. (2006). However, (4) allows for non-stationary regimes yet the overall process is stationary. Necessary conditions for m'th order stationarity of the linear model are given by Zhu et al. (2010). For the process (4) to be first order stationary, it is required that

$$0 < \sum_{k=1}^{K} p_k \psi_{k,1} < 1.$$

When $p_1 = 1$ and $p_k = 0$ for $k > 2$, then the above condition reduces to $0 < \psi_{1,1} < 1$ which is standard condition for the Poisson INARCH(1) model (see Ferland et al. (2006)) .

Other models can be easily introduced. For instance, we can consider a two regime model where in the first regime we assume that the mean process follows a linear model, such as (4), but in the second regime

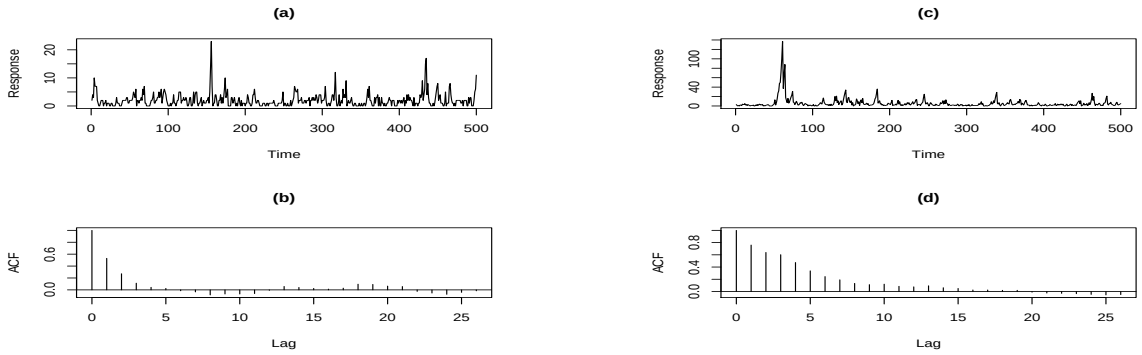

Figure 1: Left plot: Five hundred realizations obtained by model (4) where the parameter values are set to $\psi_{1,0} = 1, \psi_{1,1} = 1/4, \psi_{2,0} = 1/2$ and $\psi_{2,1} = 6/5$, with the probability of each regime being equal to 0.5. (a) Observations (b) Sample autocorrelation function. Right plot: Five hundred realizations obtained by model (7) where the parameter values are set to $\psi_{1,0} = 1, \psi_{1,1} = 1/4, \psi_{1,2} = 1/5, \psi_{2,0} = 1/2$, $\psi_{2,1} = 6/5$ and $\psi_{2,2} = 1/5$ with the probability of each regime being equal to 0.5. (c) Observations (d) Sample autocorrelation function.

a nonlinear model is employed; see for instance Fokianos and Tjøstheim (2012, Eq. (5) & (7)) and the review articles by Tjøstheim (2012, 2015). This discussion shows the great flexibility that mixture models can provide for modeling count time series.

2.2 On ergodicity and stationarity of MINARCH(∞) models

In this section, we study properties of model (3) by assuming that $L = \infty$, i.e. we deliver conditions for ergodicity and stationarity of MINARCH(∞) model. We resort to the notion of weak dependence, see Dedecker et al. (2007) for details. In particular, the appropriate notion of weak dependence for model (3) corresponds to τ–dependence as introduced by Dedecker and Prieur (2004). For a brief introduction to the τ–dependence concept see Appendix A-1. To study the properties of model (3), it is instructive to rewrite it as

$$Y_t = \sum_{k=1}^{K} \mathbb{1}_{\{Z_t=k\}} N_t(\lambda_{k,t}) = N_t(\lambda_{Z_t,t}) \tag{5}$$

where $\mathbb{1}(.)$ denotes the indicator function, Z_t is an iid sequence with $\mathrm{P}(Z_t = k) = p_k$ for $k = 1, \ldots, K$, and N_t is an iid sequence of unit rate Poisson processes which is independent of Z_t, for all t.

Let $(\Omega, \mathcal{G}, \mathrm{P})$ be a probability space and suppose that $\mathcal{M}$ is a σ-algebra of $\mathcal{G}$. We denote by $\mathcal{L}^s \equiv \mathcal{L}^s(\Omega, \mathcal{G}, \mathrm{P})$ the class of R^d-valued random variables W, such that $\|W\|_s = (E\|W\|^s)^{1/s} < \infty$. Then we obtain the following results whose proof is given in Appendix A-2.

Theorem 2.1 Consider model (3) with $L = \infty$ and assume that for any $\boldsymbol{x}$ and $\boldsymbol{x}'$ in $\mathbb{N}^\infty \times \mathbb{N}^\infty$ (with $\mathbb{N} = \{0, 1, 2, \ldots\}$), there exist sequences $(\alpha_{kj})_{j\geq 1}$, $k = 1, \ldots, K$ of non-negative real numbers such that

$$|f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}')| \leq \sum_{l=1}^{\infty} \alpha_{kl} |x_l - x'_l| \quad k = 1, 2, \ldots, K.$$

Denote by $A_k = \sum_l \alpha_{kl}$ and let $B_s = \sum_{k=1}^K p_k A_k^s < 1$. Then (i) If $s = 1$ there exists a $\tau-$weakly dependent strictly stationary process $\{Y_t, t \in \mathbb{Z}\}$ which belongs to $\mathcal{L}^1$ and (ii) for $s > 1$ this solution belongs to $\mathcal{L}^s$. Moreover

$$\tau(r) \leq \frac{2}{1 - B_1} \max_{1\leq k\leq K} f_k(\mathbf{0}) \inf_{1\leq p\leq r} \left\{ B_1^{\frac{r}{p}} + \frac{1}{1 - B_1} \sum_{q=p+1}^{\infty} \sum_{k=1}^{K} p_k \alpha_{kq} \right\}.$$

The above theorem shows that there exists a strictly stationary solution for model (3) which has finite moments and such that decay of the coefficients $\tau(\cdot)$ ensures the conditions needed for studying asymptotic inference for the maximum likelihood estimator. Furthermore, following Doukhan and Wintenberger (2008), it can be shown that the solution of (3) can be approximated by a finite order Markov stationary sequence. This is a crucial fact, since it enables approximation of infinite order model by Markov models whose order is finite, i.e. we can approximate MINARCH(∞) by MINARCH(L) for some large value of L. If $f_k(\mathbf{0}) = 0$ for $k = 1, \ldots, K$, then the stationary solution of (3) is identical to null.

Remark 2.1 In the above theorem, the constant 2 can be replaced by any other constant $C > 1/B_1$ when $B_1 > 1/2$ as this is deduced by the proof. Moreover, in the case of MINARCH(L) we can set $\alpha_{k,l} = 0$ for $l > L$ which implies a simpler, yet geometrically decreasing, upper bound for the τ coefficients; in other words

$$\tau(r) \leq \frac{2}{1 - B_1} \max_{1\leq k\leq K} f_k(\mathbf{0}) \cdot B_1^{\frac{r}{L}}.$$

2.3 The MINGARCH model

Following the works by Ferland et al. (2006) and Fokianos et al. (2009), for instance, we can extend model (3) by including a feedback process in the right hand side equation. Recall the notation introduced by (5) and let L_1 and L_2 be positive integers. Then (3) can be extended as

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t), \lambda_{k,t} = h_k(\lambda_{Z_{t-1},t-1}, \ldots, \lambda_{Z_{t-L_1},t-L_1}, Y_{t-1}, \ldots, Y_{t-L_2}), \;\; k = 1, \ldots, K. \tag{6}$$

Here the functions $h_k(\cdot)$ are non-negative and known up to a finite dimensional parameter vector and they are defined on $\mathbb{R}_+^{L_1} \times \mathbb{N}^{L_2}$ for all $k = 1, 2 \ldots, K$. We note that (6) allows dependence of the hidden process to past Z_t's in addition to past Y_t's. We will call this model as MINGARCH(L_1, L_2).

Example 2.2 Continuing Example 2.1, we have that in case that the functions $h_k(.)$ are assumed to be linear and $L_1 = L_2 = 1$, then (6) becomes

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t), \quad \lambda_{k,t} = \psi_{k,0} + \psi_{k,1} Y_{t-1} + \psi_{k,2} \lambda_{Z_{t-1},t-1}, \;\; k = 1, 2, \ldots, K. \tag{7}$$

The properties of this model are better understood by examining the right plot of Figure 1 which shows five hundred realizations and the corresponding sample autocovariance function from (7) with $K = 2$. It is clear that the inclusion of hidden process yields large correlation values which decay slower than those of model (4). Here note again that we can allow for non-stationary components in each regime yet the overall process is stationary. The same remarks made for model (4) apply in the case of model (7).

As we show next, (6) can be approximated by a MINARCH(∞) model, or in other words, it is written as

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t), \lambda_{k;t} = f_k(Y_{t-1}, Y_{t-2}, \ldots), \;\; k = 1, \ldots, K.$$

This fact is proved next for the simple case of $K = 1$ regime Poisson model. However Theorem 2.1 and (6) show that the result is true for $K > 1$.

Lemma 2.1 Suppose that $\{Y_t, t \in \mathbb{Z}\}$ is stationary count time series such that

$$Y_t = N_t(\lambda_t), \quad \lambda_t = h(\lambda_{t-1}, \ldots, \lambda_{t-L_1}, Y_{t-1}, \ldots, Y_{t-L_2})$$

where $\{N_t(\cdot), t \in \mathbb{Z}\}$ is a sequence of iid Poisson processes with unit rate and $h(\cdot)$ is a positive function. Suppose further that there exist $a_i \geq 0$, $i = 1, 2, \ldots, L_1$, $b_j \geq 0$, $j = 1, 2, \ldots, L_2$ such that

$$|h(\lambda_1, \ldots, \lambda_{L_1}, y_1, \ldots, y_{L_2}) - h(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{L_1}, \tilde{y}_1, \ldots, \tilde{y}_{L_2})| \leq \sum_{i=1}^{L_1} a_i |\lambda_i - \tilde{\lambda}_i| + \sum_{j=1}^{L_2} b_j |y_j - \tilde{y}_j|,$$

for all λ_i, $\tilde{\lambda}_i > 0$, $i = 1, \ldots, L_1$ and for all y_j, $\tilde{y}_j$ in $\mathbb{N}$, $j = 1, \ldots, L_2$ with $0 < \sum_i a_i + \sum_j b_j < 1$. Then, it holds that

$$Y_t = N_t(\lambda_t), \quad \lambda_t = f(Y_{t-1}, Y_{t-2}, \ldots)$$

where $f(.)$ is a positive function that satisfies

$$|f(y_1, y_2, \ldots) - f(\tilde{y}_1, \tilde{y}_2, \ldots)| \leq \sum_{i=}^{\infty} c_i |y_i - \tilde{y}_i|$$

and (c_i) is a sequence of positive coefficients that decay exponentially fast.

3 Maximum Likelihood Inference for MINARCH(L) models

We first investigate the behavior of maximum likelihood estimator (MLE) for model (3) when the number of regimes is known. For ordinary Poisson nonlinear models, similar studies have been given by Fokianos and Tjøstheim (2012) and Doukhan and Kengne (2015) for the MLE and by Christou and Fokianos (2014), Ahmad and Franq (2016) and Douc et al. (2017) for the Quasi-MLE. However, under the mixture model we consider in this work, some care should be taken to ensure identifiability of the parameter vector. In particular, when

the number of regimes is not known, then an identifiability issue arises and we show that this problem can be attacked by considering a penalized likelihood approach. Initially, we introduce notation and definition for developing the asymptotic theory.

Assume that the observations $\{Y_t, t = 1, \ldots, n\}$ is a realization of a strictly stationary process $\{Y_t, t \in \mathbb{Z}\}$. Denote the vector of unknown parameters by $\boldsymbol{\psi} = (\boldsymbol{\psi}_0^T, \boldsymbol{\psi}_1^T, \ldots, \boldsymbol{\psi}_K^T)^T$ where $\boldsymbol{\psi}_0^T = (p_1, \ldots, p_{K-1}) \in [0;1]^{K-1}$ and $p_K = 1 - \sum_{k=1}^{K-1} p_k \in [0;1]$. Moreover, $\boldsymbol{\psi}_k$, $k = 1, 2, \ldots, K$, denotes the parameter vector of the k'th regression function; i.e we assume that $f_k(.) := f(.; \boldsymbol{\psi}_k)$, $k = 1, 2, \ldots, K$, by recalling (3). Write d_k for the dimension of the parameter $\boldsymbol{\psi}_k$. Hence $\dim(\boldsymbol{\psi}) \equiv d = (K-1) + \sum_{k=1}^{K} d_k$. Introduce the function

$$g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi}_k) = \frac{\exp(-f(y_1, \ldots, y_L; \boldsymbol{\psi}_k)) f^y(y_1, \ldots, y_L; \boldsymbol{\psi}_k)}{y!}, \tag{8}$$

for the conditional pmf of Y_t given that it is in regime k, $k = 1, \ldots, K$. Denote by $\boldsymbol{\Psi}$ the parameter space which is assumed to be an open subset of $\mathbb{R}^d$. Finally, define the set $\mathcal{G}$ by

$$\mathcal{G} = \left\{ g : g(y, y_1, \ldots, y_L; \boldsymbol{\psi}) = \sum_{k=1}^{K} p_k g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi}_k), \boldsymbol{\psi} \in \boldsymbol{\Psi} \right\} \tag{9}$$

The MLE of the parameter vector $\boldsymbol{\psi}$ can be computed by employing the EM algorithm (Dempster et al. (1977)) as in Zhu et al. (2010). Because we are working with a simple mixture and the number of regimes is known, it is more convenient to compute directly the log-likelihood function and its derivatives, as we show next.

3.1 The log-likelihood function

Recall (5). Then, the conditional likelihood of the data for the parameter $\boldsymbol{\psi}$ is equal to

$$\begin{aligned} L(y_1, \ldots, y_n; \boldsymbol{\psi}) &= \prod_{t=L+1}^{n} L(y_t \,|y_{t-1}, \ldots, y_{t-L}\,; \boldsymbol{\psi}) \\ &= \prod_{t=L+1}^{n} \sum_{k=1}^{K} L(y_t \,|Z_t = k, y_{t-1}, \ldots, y_{t-L}\,; \boldsymbol{\psi}) \times P(Z_t = k; \boldsymbol{\psi}) \end{aligned}$$

because $P(Z_t = k; \boldsymbol{\psi}\,|\boldsymbol{y_{t-1}}, \ldots, \boldsymbol{y_{t-L}}) = P(Z_t = k; \boldsymbol{\psi}) = p_k$. By using (8) and (9), the conditional likelihood is given by

$$L(y_1, \ldots, y_n; \boldsymbol{\psi}) = \prod_{t=L+1}^{n} \Big(\sum_{k=1}^{K} p_k g_k(y_t, y_{t-1}, \ldots, y_{t-L}; \boldsymbol{\psi}_k) \Big) = \prod_{t=L+1}^{n} g(y_t, y_{t-1}, \ldots, y_{t-L}; \boldsymbol{\psi}). \tag{10}$$

Hence, we obtain that the log-likelihood function is given by

$$l_n(\boldsymbol{\psi}) = \sum_{t=L+1}^{n} \log\Big(g(y_t, y_{t-1}, \ldots, y_{t-L}; \boldsymbol{\psi}) \Big). \tag{11}$$

Note that the computation of the log-likelihood function is based on model (3) but the same calculation persists for the more general case of model (6) with obvious modifications. Turning back to model (3), the MLE is

denoted $\widehat{\boldsymbol{\psi}}_n$ and maximizes the log-likelihood function (11), that is

$$\widehat{\boldsymbol{\psi}}_n = \arg\max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} l_n\left(\boldsymbol{\psi}\right). \tag{12}$$

When the number of regimes is known, the statistical inference for autoregressive regime-switching models can be developed along the lines of standard likelihood theory. Initially, we address this simple case.

3.2 Known number of regimes

In this section we study the asymptotic behavior of the MLE, defined by (12), for the parameter $\boldsymbol{\psi}$ when the MINARCH(L) model (3) holds true. More precisely, we prove consistency and asymptotic normality. In order to study the asymptotic behavior of the MLE we assume the following standard assumptions.

H-1: The parameter vector $\boldsymbol{\psi}$ belongs to a compact subset of $\boldsymbol{\Psi}$ and the true parameter $\boldsymbol{\psi^0}$ of the model belongs to the interior of $\boldsymbol{\Psi}$.

H-2 The regression functions f_k are continuous and identifiable with respect to $\boldsymbol{\psi}_k$, i.e. $f\left(\cdot;\boldsymbol{\psi}_k\right) = f\left(\cdot;\boldsymbol{\psi}'_k\right) \Leftrightarrow \boldsymbol{\psi}_k = \boldsymbol{\psi}'_k$. In addition, they satisfy that $f_k \geq C_1$ for some constant $C_1 > 0$ and for all $k = 1,\ldots,K$.

H-3 Denote by $\partial/\partial\boldsymbol{\psi}$ the derivative with respect to all components of $\boldsymbol{\psi}$ and by

$$I_0 = -E\left(\frac{\partial^2 \log\left(g\left(Y_t,\ldots,Y_{t-L};\boldsymbol{\psi^0}\right)\right)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\right) = E\left(\frac{\partial \log\left(g\left(Y_t,\ldots,Y_{t-L};\boldsymbol{\psi^0}\right)\right)}{\partial\boldsymbol{\psi}}\frac{\partial \log\left(g\left(Y_t,\ldots,Y_{t-L};\boldsymbol{\psi^0}\right)\right)}{\partial\boldsymbol{\psi}^T}\right),$$

the Fisher information matrix, where expectation is taken with respect to the stationary distribution. Assume that the matrix I_0 exists, is invertible, and for a neighborhood $\mathcal{V}$ of $\boldsymbol{\psi^0}$:

$$E\left(\sup_{\boldsymbol{\psi}\in\mathcal{V}}\left\|\frac{\partial^2 \log\left(g\left(Y_t,\ldots,Y_{t-L};\boldsymbol{\psi^0}\right)\right)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\right\|\right) < \infty,$$

where $\|.\|$ denotes any norm in the space of $d \times d$ matrices.

H-4 The functions $f_k(\cdot)$, $k = 1,\ldots,K$ are three times differentiable with respect to $\boldsymbol{\psi}$. In addition, they satisfy for all k that

$$\begin{aligned}
\left|\frac{\partial f_k(Y_1,\ldots,Y_L;\boldsymbol{\psi})}{\partial\psi_i} - \frac{\partial f_k(Y'_1,\ldots Y'_L;\boldsymbol{\psi})}{\partial\psi_i}\right| &\leq \sum_{l=1}^{L} b_{kli}|Y_l - Y'_l|, \quad i = 1,\ldots,d,\\
\left|\frac{\partial^2 f_k(Y_1,\ldots,Y_L;\boldsymbol{\psi})}{\partial\psi_i\partial\psi_j} - \frac{\partial^2 f_k(Y'_1,\ldots,Y'_L;\boldsymbol{\psi})}{\partial\psi_i\partial\psi_j}\right| &\leq \sum_{l=1}^{L} b_{klij}|Y_l - Y'_l|, \quad i = 1,\ldots,d,\\
\left|\frac{\partial^3 f_k(Y_1,\ldots Y_L;\boldsymbol{\psi})}{\partial\psi_i\partial\psi_j\partial\psi_r} - \frac{\partial^3 f_k(Y'_1,\ldots,Y'_L;\boldsymbol{\psi})}{\partial\psi_i\partial\psi_j\partial\psi_r}\right| &\leq \sum_{l=1}^{L} b_{klijr}|Y_l - Y'_l|, \quad i = 1,\ldots,d,
\end{aligned}$$

We further assume $\forall i,j,r \in \{1,\ldots,d\}$ that $\sum_i^d b_{kli} < \infty$, $\sum_{i,j}^d b_{klij} < \infty$, $\sum_{i,j,k}^d b_{klijr} < \infty$, and $E|\partial f_k(\mathbf{0};\boldsymbol{\psi})/\partial\psi_i| < \infty$, $E|\partial^2 f_k(\mathbf{0};\boldsymbol{\psi})/\partial\psi_i\partial\psi_j| < \infty$, $E|\partial^3 f_k(\mathbf{0};\boldsymbol{\psi})/\partial\psi_i\partial\psi_j\partial\psi_k| < \infty$.

Assumptions **H-1**-**H-4** are standard in the literature and they are used to ensure a well defined model and for proving asymptotic normality of the conditional MLE. Assumption **H-1** rules out the possibility of the true parameter to belong to the boundary of the parameter space. Condition **H-2** is equivalent to $\boldsymbol{\psi}_0$ being a locally unique asymptotic maximizer of the log–likelihood function (10); see Berkes et al. (2003, Theorem 2.3) and Francq and Zakoïan (2004, Assumption A4 and Remark 2.4). Assumption **H-3** is implied by assuming that the elements of the vector $\partial \log(g(y_{t-1}, \ldots, y_{t-L}))/\partial \boldsymbol{\psi}$ are linearly independent. Finally, **H-4** implies that the functions $f_k(\cdot)$ are sufficiently smooth for all k, so that higher order derivatives of the conditional log-likelihood function exist and are finite. Before proceeding to the main results we show that if the number of regimes is known, then the model (4) is identifiable up to a permutation. The following lemma is proved in the appendix.

Lemma 3.1 Consider model (4). Then, under assumption **H-2**, there exists a permutation σ such that for all $(y, y_1, \ldots, y_L)$

$$\sum_{k=1}^{K} p_k g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi}_k) = \sum_{k=1}^{K} p'_k g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi}'_k) \Longleftrightarrow \sigma(\boldsymbol{\psi}) = \boldsymbol{\psi}',$$

where $\sigma(\boldsymbol{\psi}) = (\sigma(\boldsymbol{\psi}_0^T), \boldsymbol{\psi}_{\sigma(1)}^T, \ldots, \boldsymbol{\psi}_{\sigma(K)}^T)^T$ and $\sigma(\boldsymbol{\psi}_0) = (p_{\sigma(1)}, \ldots, p_{\sigma(K-1)})^T$.

The next result is a consequence of the results obtained in previous mentioned references. We will not prove this in detail but in the Appendix we outline computations of the score function and information matrix. The proof follows by employing Taniguchi and Kakizawa (2000, Thm 3.2.23).

Theorem 3.1 Consider model (3) and suppose that assumptions **H-1**–**H-4** and the conditions of Theorem 2.1, for $s \geq 4$, hold true. Then, there exists a fixed open neighborhood $O(\boldsymbol{\psi}^0)$ of $\boldsymbol{\psi}^0$ such that with probability tending to 1 as $n \to \infty$, the equation

$$\frac{\partial l_n(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \sum_{t=1}^{n} \frac{\partial \log(g(y_t, y_{t-1}, \ldots, y_{t-L}); \boldsymbol{\psi}^0)}{\partial \boldsymbol{\psi}^0} = 0$$

has a unique solution which is denoted by $\widehat{\boldsymbol{\psi}}_n$ as in (12). Furthermore, the MLE estimator is strongly consistent, i.e

$$\widehat{\boldsymbol{\psi}}_n \xrightarrow{a.s.} \boldsymbol{\psi}^{\mathbf{0}},$$

and asymptotically normally distributed

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^{\mathbf{0}}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I_0^{-1}\right),$$

as $n \to \infty$.

3.3 Unknown number of regimes

When the number of regimes is unknown, the identifiability problem cause the likelihood ratio test statistic (LRT) not to converge to a χ^2-distribution, which is a consequence of standard theory. A simple example

illustrates the identifiability problem. Assume a linear model for (3) with $L = 1$ and suppose that we want estimate a model with $K = 2$ regimes. However, the true model has only $K^0 = 1$ regime i.e.

$$P(Y_t = y \,|Y_{t-1}\,) = \exp(-(\psi_0^0 + \psi_1^0 Y_{t-1})\frac{(\psi_0^0 + \psi_1^0 Y_{t-1})^y}{y!}.$$

A linear model with $K = 2$ regimes has conditional pmf

$$P(Y_t = y \,|Y_{t-1}\,) = p\exp(-(\psi_{1,0} + \psi_{1,1} Y_{t-1}))\frac{(\psi_{1,0}+\psi_{1,1}Y_{t-1})^y}{y!} + (1-p)\exp(-(\psi_{2,0} + \psi_{2,1} Y_{t-1}))\frac{(\psi_{2,0}+\psi_{2,1}Y_{t-1})^y}{y!}.$$

Hence, any parameter vector $\boldsymbol{\psi} = (p, \psi_{1,0}, \psi_{1,1}, \psi_{2,0}, \psi_{2,1})$ such that $\psi_{1,0} = \psi_{2,0} = \psi_0^0, \psi_{1,1} = \psi_{2,1} = \psi_1^0,\ p \in [0;1]$ or $\psi_{2,0} = \psi_0^0, \psi_{2,1} = \psi_1^0, (\psi_{1,0}, \psi_{1,1}) \in \mathbb{R}^2,\ p = 0$ or $\psi_{1,0} = \psi_0^0, \psi_{1,1} = \psi_1^0, (\psi_{2,0}, \psi_{2,1}) \in \mathbb{R}^2,\ p = 1$ satisfies to the true conditional pmf. This implies that the Fisher information matrix is not invertible for a model whose number of regimes is overestimated. However, we show next that we can still obtain the asymptotic distribution of the LRT. Following recent work by Olteanu and Rynkiewicz (2012), we compute the asymptotic distribution of the LRT, under suitable conditions, in the case that the number of regimes is overestimated.

Suppose that K^0 denotes the true number of regimes. If $K \leq K^0$ there are no identification issues with model (3); therefore we assume that $K > K^0$ in the sequel. The LRT for testing the hypothesis $\mathrm{H}_0 : K = K^0$ is given by

$$2\lambda_n = 2\left(\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \ln(\boldsymbol{\psi}) - \ln(\boldsymbol{\psi^0})\right) = 2\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \frac{\sum_{t=P}^{n}\sum_{k=1}^{K} p_k g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi})}{\sum_{t=P}^{n}\sum_{k=1}^{K^0} p_k^0 g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi^0})} \tag{13}$$

where $\boldsymbol{\psi^0}$ is a parameter value satisfying the true density function g^0:

$$g^0(y, y_1, \ldots, y_L) = \sum_{k=1}^{K^0} p_k^0 g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi^0})$$

There might exist an infinite number of such parameters but the true mass function is unique.

To prove the main result in this section we will use Doukhan et al. (2012, Thm.2) which states that every τ-weakly dependent multivariate integer valued stationary Markov chain is also β–mixing (for appropriate definitions see Doukhan (1994)). We will denote the β-mixing coefficients by (β_r), $r \in \mathbb{N}$. Recall (3). Then, the $(L+1)$-dimensional process $(Y_t, \ldots, Y_{t-L})^T$ is Markov chain taking values in the discrete state space $\mathbb{Z}^{L+1}$. Hence, Thm. 2 of Doukhan et al. (2012) applies and shows that the process is is geometrically β-mixing process. This results has been proved by Neumann (2011) for the case $L = 1$ and and some more recent related work has been given by Truquet (2017) and Doukhan and Neumann (2017). Denote by μ the corresponding stationary measure of the vector $(Y_t, Y_{t-1}, \ldots, Y_{t-L})$. We will also need the following notation. For $\eta > 0$, denote by $\mathcal{G}_\eta = \left\{g \in \mathcal{G}, \|g - g^0\|_{L^2(\mu)} \leq \eta\right\}$. The extended set of score-functions $\mathcal{S}_\eta$ is defined as:

$$\mathcal{S}_\eta = \left\{ s_g = \frac{\frac{g}{g^0} - 1}{\left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)}}, g \in \mathcal{G}_\eta \right\}.$$

We also define the limit-set of scores $\mathcal{D}$

$$\mathcal{D} = \left\{ d \in \mathbb{L}^2(\mu) \mid \exists\ (g_n) \in \mathcal{G},\ \| \frac{g_n - g^0}{g^0} \|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0, \| d - s_{g_n} \|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0 \right\}.$$

Setting $g_t = g_n$ for $t \in [0,1]$ and $n \leq \frac{1}{t} < n+1$, we obtain that, for all $d \in \mathcal{D}$, there exists a parametric path $(g_t)_{0\leq t\leq 1}$ such that $\forall\ t \in [0,1]$, $g_t \in \mathcal{G}$, $t \to \| \frac{g_t - g^0}{g^0} \|_{\mathbb{L}^2(\mu)}$ is continuous, $\| \frac{g_t - g^0}{g^0} \|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$ and $\| d - s_{g_t} \|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$. With the previous notations, we introduce the following assumption

T-1 The set $\mathcal{G}$ is Glivenko-Cantelli and the parameter space $\mathbf{\Psi}$ contains a neighborhood of the parameters defining the true conditional density g^0.

T-2 There exists $\eta > 0$ such that for all $g \in \mathcal{G}$ with $\| g - g^0 \|_{L^2(\mu)} \leq \eta$, $\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)} < \infty$

Following Olteanu and Rynkiewicz (2012), we have the following theorem

Theorem 3.2 Assume **T-1**–**T-2** and **H-4** and let the conditions of Theorem 2.1 be true for $s \geq 4$. Then there exists a centered Gaussian process $\{W_S, S \in \mathbb{F}\}$ with continuous sample path and covariance kernel $P(W_{S_1} W_{S_2}) = P(S_1 S_2)$ such that

$$\lim_{n\to\infty} 2\lambda_n = \sup_{S\in\mathbb{F}} \left(\max(W_S, 0) \right)^2,$$

where $2\lambda_n$ is defined by (13). The index set $\mathbb{F}$ is defined as $\mathbb{F} = \cup_t \mathbb{F}_t$, with the union taken over $t = (t_0, \ldots, t_{K^0}) \in \mathbb{N}^{K^0+1}$ with $0 = t_0 < t_1 < \cdots < t_{K^0} \leq K$ and

$$\mathbb{F}_t = \left\{ \Omega \left(\sum_{k=1}^{K^0} \zeta_k l_{\boldsymbol{\psi}^0{}_k} + \sum_{k=K^0+1}^{p} \zeta_k l_{\boldsymbol{\psi}_k} + \sum_{k=1}^{K^0} \lambda_k^T l'_k + \delta \sum_{k=1}^{K^0} \sum_{l=t_{i-1}+1}^{t_i} \gamma_l^T l''_k \gamma_l \right), \lambda_1, \ldots, \lambda_{K^0}, \gamma_1, \ldots, \gamma_{t_{K^0}} \in \mathbb{R}^P; \right.$$
$$\left. \zeta_1, \ldots, \zeta_K \in \mathbb{R},\ \boldsymbol{\psi}_{t_{K^0}+1}, \ldots, \boldsymbol{\psi}_K \in \mathbf{\Psi} - \left\{ \boldsymbol{\psi}^0{}_1, \ldots, \boldsymbol{\psi}^0{}_{K^0} \right\} \right\}$$

where $\delta = 1$ if there exists a vector $\mathbf{q}$ such that $q_l \leq 0$, $\sum_{l=t_{k-1}+1}^{t_k} q_l = 1$, $\sum_{l=t_{k-1}+1}^{t_k} \sqrt{q_l}\gamma_l = 0$ for $k = 1, \ldots, K^0$; and $\delta = 0$, otherwise. In addition, we denote by $l'_k = \partial l_{\boldsymbol{\psi}_k} / \partial \boldsymbol{\psi}_k \left(\boldsymbol{\psi}_k^0 \right)$, and $l''_j = \partial^2 l_{\boldsymbol{\psi}_k} / \partial \boldsymbol{\psi}_k{}^2 \left(\boldsymbol{\psi}_k^0 \right)$.

For the special case of model (4) we will see in Sec. A-5 of the Appendix that both $\mathbf{T1} - \mathbf{T2}$ are satisfied. But the proof shows that it can be extended to include the case of order L linear models. The previous theorem shows that asymptotic law of the LRT depends on the true parameters of the model. The next result illustrates the implications of this theorem. For $l \in \mathbb{N}_+$, let us denote

$$\mathcal{G}_l = \left\{ g(y, y_1, \ldots, y_L; \boldsymbol{\psi}) = \sum_{k=1}^{l} p_k g_k(y, y_1, \ldots, y_L; \boldsymbol{\psi}), \boldsymbol{\psi} \in \mathbf{\Psi} \right\}.$$

For some fixed $L \in \mathbb{N}_+$ sufficiently large, we shall consider the following class of functions

$$\mathcal{G}_L = \bigcup_{l=1}^{L} \mathcal{G}_l$$

For every $g \in \mathcal{G}_L$ define the number of regimes as

$$l(g) = \min\{l \in \{1, \ldots, L\}, g \in \mathcal{G}_l\}.$$

Then, $l_0 = l\left(g^0\right)$ denotes the number of regimes of the true model. An estimate of the number of regimes, say $\hat{l}$ is defined as the integer $l \in \{1, \ldots, L\}$ which maximizes the following the penalized criterion,

$$T_n(l) = \sup_{g \in \mathcal{G}_l} l_n(g) - \alpha_n(l), \tag{14}$$

where $l_n(g)$ is given by (11) and $\alpha_n(\cdot)$ is a suitably chosen sequence.

Proposition 3.1 Assume **T-1**–**T-2** and **H-4** and let the conditions of Theorem 2.1 be true for $s \geq 4$. Let $(\alpha_n(\cdot))$ be an increasing function of l such that $a_n(l_1) - a_n(l_2) \xrightarrow[n\to\infty]{} \infty$ for every $l_1 > l_2$ and $a_n(l)/n \xrightarrow[n\to\infty]{} 0$ for every l. Then, as $n \to \infty$, the estimator $\hat{l}$, defined by maximizing (14), converges in probability to the true number of regimes, i.e.

$$\hat{l} \xrightarrow[n\to\infty]{\mathbb{P}} K^0$$

.

The proof of the above results is based on arguments given for proving Olteanu and Rynkiewicz (2012, Cor. 2.3) and therefore it is omitted. Penalization like the Bayesian information criterion (BIC) fulfills the assumptions of Proposition 3.1 and it gives a theoretical justification of the empirical results obtained by Zhu et al. (2010) for the case of linear model. Proposition 3.1 applies to nonlinear models though.

4 On likelihood inference for the MINGARCH model

Recall now model (6). Then the likelihood equations (11) are still true for this case but the complexity of computation required to fit (6) is of exponential order. Indeed, consider the simplest case of a linear model

$$\lambda_{t,k}(\boldsymbol{\psi}_k) = d_k + \sum_{i=1}^{L_1} \psi_{i,k} \lambda_{Z_{t-i},t-i} + \sum_{j=1}^{L_2} \psi_{j,k} Y_{t-j} \quad k = 1, \ldots, K.$$

For calculating the log-likelihood function, we see that we need to consider all the possible states for $Z_1, \cdots, Z_n$ and therefore the complexity of such computation is of the order K^n. Hence, in general, likelihood inference is intractable unless the number of observations is very small. Furthermore, note that knowledge of the expected value of $\lambda_{Z_{t-1},t-1}, \cdots, \lambda_{Z_{t-q},t-q}$ is not enough to compute the conditional distribution of $\lambda_{t,1}(\boldsymbol{\psi}_1), \cdots, \lambda_{t,K}(\boldsymbol{\psi}_K)$ and therefore an EM algorithm cannot be applied for likelihood optimization.

An alternative way to deal with these issues could be to employ Lemma 2.1 and use the infinite representation MINARCH(∞) of (6) with $f_k(\cdot)$ being linear. Then using a finite order would suffice to make likelihood inference for such models. However, the following counterexample proves that misspecification of the model doesn't allow anymore to estimate the true number of regimes.

Example 4.1 Assume that the observations are a realization of a stationary INGARCH(1,1) process $(Y_k)_{k\in\mathbb{N}}$, vis.

$$Y_t \mid \mathcal{F}_{t-1} \sim \mathrm{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t) \quad \lambda_t = \psi_0^0 + \psi_1^0 Y_{t-1} + \psi_2^0 \lambda_{t-1}.$$

Here, we assume that the process has only one regime. The practitioner does not know the true data generating process and fits a MINARCH(1) model with a 2 regimes. In this case, the vector of unknown parameters is $\boldsymbol{\psi} = (p_1, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2)^T$. With this notation, we obtain by (10) that the conditional likelihood is

$$\begin{aligned} L(y_1, \ldots, y_n; \boldsymbol{\psi}) &= \prod_{t=1}^{n} \Big(\sum_{k=1}^{2} p_k \frac{\exp(-\lambda_{t,k}(\boldsymbol{\psi}_k)) \lambda_{t,k}^{y_t}(\boldsymbol{\psi}_k)}{y_t!} \Big) \\ &= \prod_{t=1}^{n} \Big(p_1 \frac{\exp(-(\psi_{1,0} + \psi_{1,1} y_{t-1}))\, (\psi_{1,0} + \psi_{1,1} y_{t-1})^{y_t}}{y_t!} \\ &\quad + (1-p_1) \frac{\exp(-(\psi_{2,0} + \psi_{2,1} y_{t-1}))\, (\psi_{2,0} + \psi_{2,1} y_{t-1})^{y_t}}{y_t!} \Big). \end{aligned}$$

In what follows expectation is taken wrt to stationary measure of (Y_1, Y_2) which exists by Theorem 2.1. The negative of the expected log-likelihood converges to the Kullback-Leibler distance (up to a constant), by ergodicity. In other words

$$\begin{aligned} -\frac{\log L(y_1, \cdots, y_n; \boldsymbol{\psi})}{n} \xrightarrow{a.s.} K(\boldsymbol{\psi}) \doteq \mathrm{E}\Big[\log\Big(& p_1 \frac{\exp(-(\psi_{1,0} + \psi_{1,1} Y_1))\, (\psi_{1,0} \psi_{1,1} Y_1)^{Y_2}}{Y_2!} \\ & + (1-p_1) \frac{\exp(-(\psi_{2,0} + \psi_{2,1} Y_1))\, (\psi_{2,0} + \psi_{2,1} Y_1)^{Y_2}}{Y_2!} \Big) \Big]. \end{aligned}$$

The "best" model (which is not the true model) is the model that minimizes the Kullback-Leibler distance

$$K(\boldsymbol{\psi}^*) = \arg\min_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} K(\boldsymbol{\psi}).$$

The problem that we are faced with is to examine whether the best model has one or two regimes. If the best model has only one regime then $\boldsymbol{\psi}^* = \left(p_1^*, \psi_{1,0}^*, \psi_{1,1}^*, \psi_{2,0}^*, \psi_{2,1}^* \right)$ has to satisfy $p_1^* = 0$, or $p_1^* = 1$ or $(\psi_{1,0}^*, \psi_{1,1}^*) = (\psi_{2,0}^*, \psi_{2,1}^*)$. If these conditions are not satisfied, then the "best" model will have two regimes but the true number of regimes is only one.

Hence the vector $\boldsymbol{\psi}^*$ has to satisfy the following equations

$$-\mathrm{E}\Big[\frac{\partial}{\partial \psi_i^*} \log \left(p_1 \frac{\exp(-(\psi_{1,0} + \psi_{1,1} Y_1))\, (\psi_{1,0} + \psi_{1,1} Y_1)^{Y_2}}{Y_2!} + (1-p_1) \frac{\exp(-(\psi_{2,0} + \psi_{2,1} Y_1))\, (\psi_{2,0} + \psi_{2,1} Y_1)^{Y_2}}{Y_2!} \right) \Big] = 0,$$

where ψ_i^* is the i'th component of $\boldsymbol{\psi}^*$, $i = 1, \ldots, 5$. By taking into account the calculations in Appendix A-5 we

have that if $p_1^* = 0$, then the above system reduces to

$$\begin{aligned} -\mathrm{E}\left[\frac{\exp(-(\psi_{1,0}+\psi_{1,1}Y_1))\,(\psi_{1,0}+\psi_{1,1}Y_1)^{Y_2}}{\exp(-(\psi_{2,0}+\psi_{2,1}Y_1))\,(\psi_{2,0}+\psi_{2,1}Y_1)^{Y_2}}-1\right] &= 0, \\ -\mathrm{E}\left[\left(\frac{Y_2}{\psi_{2,0}+\psi_{2,1}Y_1}-1\right)\right] &= 0, \\ -\mathrm{E}\left[\left(\frac{Y_1Y_2}{\psi_{2,0}+\psi_{2,1}Y_1}-1\right)\right] &= 0. \end{aligned}$$

The last equation shows that

$$\mathrm{E}\left[\left(\frac{Y_1Y_2}{\psi_{2,0}+\psi_{2,1}Y_1}-1\right)\right] = \mathrm{E}\left[\mathrm{E}\left[\left(\frac{Y_1Y_2}{\psi_{2,0}+\psi_{2,1}Y_1}-1\right)|Y_1\right]\right] = E_\mu\left[Y_1-1\right] \neq 0.$$

By a symmetric argument, if $p_1^* = 1$ we obtain that $\mathrm{E}\left[Y_2-1\right] = 0$ which is false again. Finally, for the case of $(\psi_{1,0},\psi_{1,1}) = (\psi_{2,0},\psi_{2,1})$, we get the following system

$$\begin{aligned} -\mathrm{E}\left[p_1\left(\frac{Y_2}{\psi_{1,0}+\psi_{1,1}Y_1}-1\right)\right] &= 0, \\ -\mathrm{E}\left[p_1\left(\frac{Y_1Y_2}{\psi_{1,0}+\psi_{1,1}Y_1}-1\right)\right] &= 0, \\ -\mathrm{E}\left[(1-p_1)\left(\frac{Y_2}{\psi_{1,0}+\psi_{1,1}Y_1}-1\right)\right] &= 0, \\ -\mathrm{E}\left[(1-p_1)\left(\frac{Y_1Y_2}{\psi_{1,0}+\psi_{1,1}Y_1}-1\right)\right] &= 0. \end{aligned}$$

Arguing as before, we have that this system of equations is again false, in general. Hence we illustrated that approximating a MINGARCH model with an MINARCH model will not allow to estimate the true number of regimes. It is easy to see that his problem will occur even if we had applied MINARCH models with lag L bigger than 1.

5 A Data Example

In this section, we apply the methodology to the time series shown at the top of Figure 2 which consists of weekly number of disease cases caused by E.coli in the state of North Rhine-Westphalia (Germany) from January 2001 to May 2013. The data are available in the R package `tscount`; Liboschik et al. (2017). As a first remark we note that there was an outbreak of E.coli infections all over Northern Germany around middle May to end of June 2011; for further details see `https://en.wikipedia.org/wiki/2011_Germany_E._coli_O104:H4_outbreak`. The plot reveals the outbreak quite clearly and it suggests two possible modeling approaches. The one is the approach taken in this work since the graph illustrates the existence of one region with low activity and another region of higher activity. A second possible approach for analysing these data is that

of Fokianos and Fried (2010) who consider various types of interventions and develop methodology for their detection. In what follows we will discuss both approaches. The time series of E.coli cases consist of 646 observations but we removed the first three observations for model fitting; hence $n = 643$ in what follows.

To fit model (4) we employ the transformation $p_k = \exp(\theta_k)/(1 + \sum_{j=1}^{K} \exp(\theta_j))$, $k = 1, \ldots, K$ and $\theta_K = 0$ to avoid numerical instability. Table 1 shows the values of BIC after fitting a model with $K = 1$ (ordinary Poisson model) and $K = 2$ components model for various choices of lag L. The BIC favors a two-component model where in each region Y_t is regressed on Y_{t-1}, Y_{t-2}. The parameter estimates and their standard errors for $\boldsymbol{\psi} = (\theta_1, \psi_{1,0}, \psi_{1,1}, \psi_{1,2}, \psi_{2,0}, \psi_{2,1}, \psi_{2,2})^T$ are $(-0.597, 9.475, 0.573, 0.262, 5.431, 0.344, 0.226)^T$ and $(0.068, 1.123, 0.003, 0.003, 0.641, 0.001, 0.002)^T$. These results can be interpreted broadly as follows. The probability of one regime is $p_1 = 0.36$ and the probability of the other regime is $p_2 = 1 - p_1 = 0.64$. In the region which has the smaller probability, the mean number of weakly cases is larger than the mean number of weakly cases in the region with higher probability. This interpretation agrees with Figure 2 where the plots reveals this specific structure and indicates some correlation of the response to Y_{t-1} and Y_{t-2}.

It is worth comparing this approach with the methodology developed by Fokianos and Fried (2010) for detecting intervention in count time series and is implemented in the R package `tscount`. Indeed, some further data analysis shows the existence of a transient effect at time $t = 540$ which corresponds to the week starting at May 23rd, 2011. In this case, the fitted model is

$$\begin{aligned} Y_t \mid \mathcal{F}_{t-1} &\sim \text{Poisson}(\lambda_t) \\ \lambda_t &= 8.478(.533) + 0.339(.027)Y_{t-1} + 0.219(.026)Y_{t-2} + 33.120(3.51)\frac{1}{(1-0.9\mathcal{B})}\mathbb{1}_t(540) \end{aligned}$$

where $\mathcal{B}$ is the shift operator such that $\mathcal{B}^i Y_t = Y_{t-i}$ and $I_t(\tau)$ is an indicator function, with $I_t(\tau) = 1$ if $t = \tau$, and $I_t(\tau) = 0$ if $t \neq \tau$. Corresponding standard errors of the regression coefficients are in parentheses. The corresponding BIC values obtained after fitting this model is equal to 4411.754 which improved the model without intervention; see Table 1 for $K = 1$ and $L = 2$. Yet, it seems that a mixture model with two regimes might be a better alternative for modeling these data. This is also supported informally by considering the Pearson residuals defined by

$$e_t = \frac{Y_t - \mathrm{E}[Y_t \mid \mathcal{F}_{t-1}]}{\sqrt{\mathrm{Var}[Y_t \mid \mathcal{F}_{t-1}]}}$$

In the case of model (4), use (2) to calculate the above quantity. The mean square error of the Pearson residuals is equal to 1.998 when using the model with intervention. Model (4) has mean square error equal to 1.13. So there is some further evidence supporting the mixture approach for modeling these data. Finally, the top graph of Figure 3 shows plots of the autocorrelation function of the Pearson residuals after fitting both models discussed to the data. In both cases, the plot does not indicate any gross departure from white noise but it should be used cautiously in practice. The lower plot of Figure 3 shows plot of the auto-distance correlation function. This quantity quantifies dependence better as it was explained by Székely et al. (2007) (for the

independent case) and Fokianos and Pitsillou (2018) for the time series case. Both of the plots indicate that the fit of both models to weakly number of E.coli cases is satisfactory.

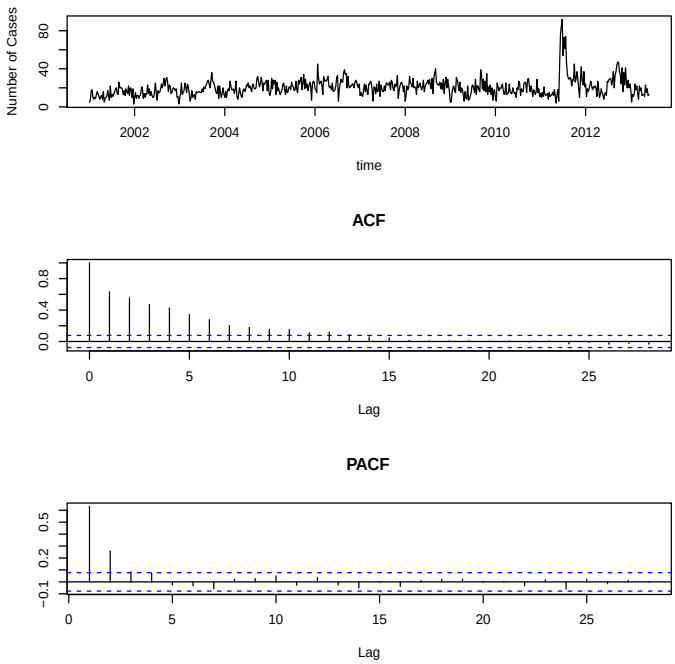

Figure 2: Weekly number of reported disease cases caused by E.coli in the state of North Rhine-Westphalia (Germany) from January 2001 to May 2013 and their sample autocorrelation and partial autocorrelation functions.

L	$K = 1$	$K = 2$
1	4636.327	4364.537
2	4540.943	**4319.091**
3	4522.201	4328.674

Table 1: BIC values after fitting model (4) to E.coli count time series.

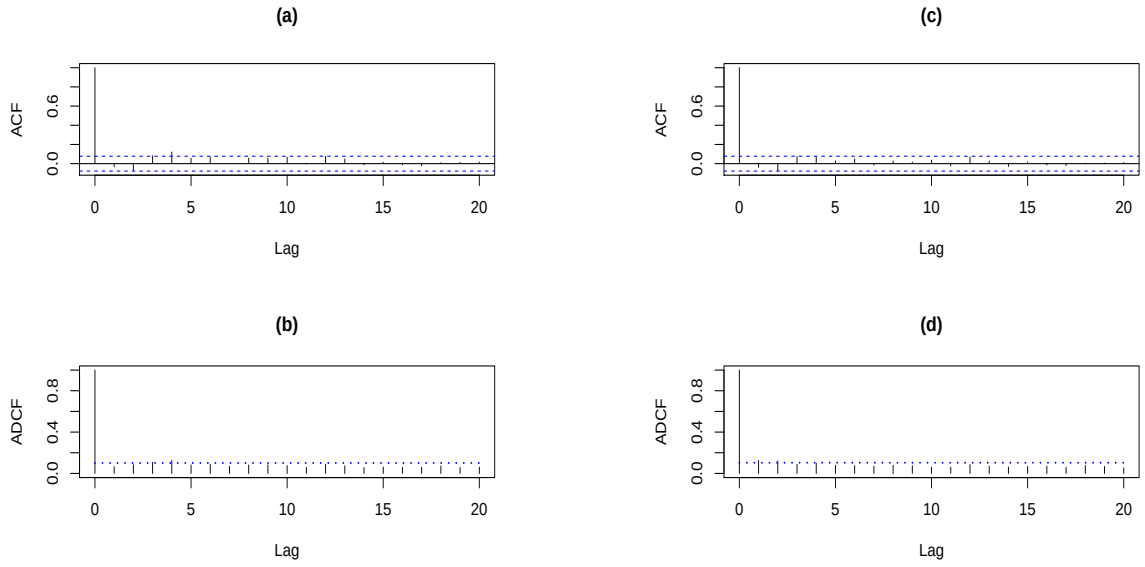

Figure 3: Left plot: (a) Autocorrelation and (b) Auto-Distance Correlation function of the Pearson residuals after fitting the model with intervention to the weekly number of E.coli cases. Right plot: (c) Autocorrelation and (d) Auto-Distance Correlation function of the Pearson residuals after fitting (4) with $K = 2$ and $L = 2$ to the weekly number of E.coli cases.

Acknowledgements

Part of this research was carried out while K. Fokianos was with the Department of Mathematics & Statistics, University of Cyprus. K. Fokianos 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.

Appendix

The following auxiliary lemma can be proved easily upon recalling the s-moments of a Poisson random variable, see Johnson et al. (1992) and Ferland et al. (2006).

Lemma A-1 Recall (1) and let $Y \sim \mathrm{MP}(\boldsymbol{p}, \boldsymbol{\lambda})$ and $s \in \mathbb{N}^*$. The uncentered moments of Y satisfy

$$\mathrm{E}Y^s = \sum_{k=1}^{K} p_k \sum_{j=0}^{s} \left\{ \begin{array}{c} s \\ j \end{array} \right\} \lambda_k^j,$$

where $\left\{ \begin{array}{c} s \\ j \end{array} \right\}$ denote Stirling numbers of the second kind such that when $s \geq 0$ and $0 \leq j \leq s$, $\left\{ \begin{array}{c} s \\ j \end{array} \right\} = 0$ for $j \notin \{1, \ldots, s\}$ and satisfy the following recurrence:

$$\left\{ \begin{array}{c} s \\ j \end{array} \right\} = \left\{ \begin{array}{c} s-1 \\ j-1 \end{array} \right\} + j \left\{ \begin{array}{c} s-1 \\ j \end{array} \right\}$$

A-1 On τ-dependennce

For the Euclidean space $\mathbb{R}^d$ and $h : \mathbb{R}^d \to \mathbb{R}$, we denote by $\|h\|_\infty = \sup_{x \in \mathbb{R}^d} |h(x)|$. In addition, let

$$\mathrm{Lip}(h) = \sup_{x \neq y} \frac{|h(x) - h(y)|}{\|x - y\|}.$$

The space $\Lambda_1(\mathbb{R}^d)$ is the set of functions $h : \mathbb{R}^d \to R$ such that $\mathrm{Lip}(h) \leq 1$. Let $(\Omega, \mathcal{G}, \mathrm{P})$ be a probability space and suppose that $\mathcal{M}$ is a σ-algebra of $\mathcal{G}$. We denote by $\mathcal{L}^s \equiv \mathcal{L}^s(\Omega, \mathcal{G}, \mathrm{P})$ the class of R^d-valued random variables W, such that $\|W\|_s = (E\|W\|^s)^{1/s} < \infty$. Let X be a random variable with values in $\mathbb{R}^d$. Assume that $\|X\|_1 < \infty$ and define the coefficient τ as

$$\tau(\mathcal{M}, X) = \left\| \sup \left\{ \left| \int f(x) \mathrm{P}_{X|\mathcal{M}}(dx) - \int f(x) \mathrm{P}_X(dx) \right| \Big/ f \in \Lambda_1(E) \right\} \right\|_1.$$

An easy way to bound this coefficient is based on a coupling argument which shows that

$$\tau(\mathcal{M}, X) \leq \|X - Y\|_1,$$

for any random variable Y with the same distribution as X and independent of $\mathcal{M}$, see Dedecker and Prieur (2004). Moreover, if the probability space $(\Omega, \mathcal{G}, \mathrm{P})$ is rich enough, then there exists such an X^* with $\tau(\mathcal{M}, X) = \|X - X^*\|_1$. Using the definition of τ, the dependence between the past of the sequence $(X_t)_{t\in Z}$ and its future k-tuples may be assessed as follows. Consider the norm $\|x - y\| = \|x_1 - y_1\| + \cdots + \|x_k - y_k\|$ on $\mathbb{R}^{dk}$, set $\mathcal{M}_p = \sigma(X_t, t \leq p)$ and

$$\begin{aligned} \tau_k(r) &= \max_{1\leq l\leq k} \frac{1}{l} \sup\Big\{\tau(\mathcal{M}_p, (X_{j_1}, \ldots, X_{j_l}))\Big/ p + r \leq j_1 < \cdots < j_l\Big\}, \\ \tau_\infty(r) &= \sup_{k>0} \tau_k(r). \end{aligned}$$

For the sake of simplicity $\tau_\infty(r)$ is denoted by $\tau(r)$. The time series $(X_t)_{t\in Z}$ is τ-weakly dependent when $\lim_{r\to\infty} \tau(r) = 0$.

A-2 Proof of Theorem 2.1

Recall that $A_k = \sum_l \alpha_{kl}$ and $B_s = \sum_{k=1}^K p_k A_k^s$. We will show that the following statements hold true:

1. If $B_1 = \sum_{k=1}^K p_k A_k < 1 < 1$, then there exists a weakly dependent strictly stationary process $\{Y_t, t \in \mathbb{Z}\}$ which belongs to $\mathcal{L}^1$.
2. If $B_s < 1$, then the strictly stationary process $\{Y_t, t \in \mathbb{Z}\}$ belongs to $\mathcal{L}^s$, for $s \in \mathbb{N}^*$.
3. If $B_s < 1$, then the strictly stationary process $\{Y_t, t \in \mathbb{Z}\}$ belongs to $\mathcal{L}^s$, for $s \in [1, \infty)$.

Step 1. The proof of point 1 is based on verifying condition (3.1) of Doukhan and Wintenberger (2008) since their condition (3.2) is assumed while condition (3.3) in the same paper trivially holds. We denote by MP_t is a sequence of iid mixture homogeneous Poisson processes with $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $\boldsymbol{\lambda} = (1, \ldots, 1)$ and by model definition we have that $Y_t = \mathrm{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t)$, where

$$\boldsymbol{\lambda}_t = (\lambda_{t;1}, \ldots, \lambda_{t;k}) = (f_1(\boldsymbol{\lambda}_{t-1}, Y_{t-1}), \ldots, f_k(\boldsymbol{\lambda}_{t-1}, Y_{t-1})) = \boldsymbol{f}(\boldsymbol{\lambda}_{t-1}, Y_{t-1}),$$

where we have defined the K-dimensional function $\boldsymbol{f} = (f_1, \ldots, f_k)$. For this model, the noise sequence can be written as $\mathrm{MP} = (Z, N)$ for a random variable Z with $\boldsymbol{P}(Z = k) = p_k$ and N a unit rate Poisson process independent of Z. Note that (recall (5))

$$\mathrm{MP}(\boldsymbol{p}, \boldsymbol{f}(\boldsymbol{x})) = \sum_{k=1}^K \mathbb{1}_{\{Z=k\}} N(f_k(\boldsymbol{x})).$$

It is seen that (3) is expressed as $Y_t = \mathrm{MP}(\boldsymbol{p}, \boldsymbol{\lambda}_t) = F(Y_{t-1}, \ldots, \mathrm{MP}_t)$. Then, with $\boldsymbol{x} = (x_1, x_2, \ldots) \in N^\infty$ and $\boldsymbol{x}'$ defined analogously, condition (3.1) of Doukhan and Wintenberger (2008) becomes

$$\begin{aligned} \mathrm{E}|F(\boldsymbol{x}, \mathrm{MP}) - F(\boldsymbol{x}', \mathrm{MP})| &= \mathrm{E}\Big|\mathrm{MP}(\boldsymbol{f}(\boldsymbol{x})) - \mathrm{MP}(\boldsymbol{f}(\boldsymbol{x}'))\Big| \\ &= \sum_{k=1}^{K} p_k |f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}')| \\ &\leq \sum_{k=1}^{K} p_k \sum_{l=1}^{\infty} \alpha_{kl} |x_l - x_l'|, \end{aligned}$$

where the second equality follows from the properties of the Poisson process. In addition $\mathrm{E}|F(\mathbf{0}, \mathrm{MP})| = \sum_{k=1}^{K} p_k f_k(\mathbf{0})$. Hence, the first part of the theorem holds.

Step 2. To obtain the second part of the theorem, we use recursion. We give a one step recursion for the special case of $s \in \mathbb{N}$. For $k = 1, \ldots, K$, we have that $f_k(\boldsymbol{x}) = f_k(\boldsymbol{x}) - f_k(\mathbf{0}) + f_k(\mathbf{0})$ thus the assumption of the theorem implies with $c_k \equiv f_k(\mathbf{0})$ that

$$|f_k(\boldsymbol{x})| \leq g_k(\boldsymbol{x}) + c_k, \qquad g_k(\boldsymbol{x}) = \sum_{l=1}^{\infty} \alpha_{kl} |x_l|.$$

Therefore we obtain that

$$|f_k(\boldsymbol{x})|^i \leq \sum_{j=0}^{i} \binom{i}{j} c_k^{i-j} g_k^j(\boldsymbol{x}) \quad \text{for } i < s$$

and

$$|f_k(\boldsymbol{x})|^s \leq g_k^s(\boldsymbol{x}) + \sum_{j=0}^{s-1} \binom{s}{j} c_k^{s-j} g_k^j(\boldsymbol{x}) \leq g_k^s(\boldsymbol{x}) + r_k(\boldsymbol{x}). \tag{A-1}$$

Now Jensen's inequality with probability weights a_{kl}/A_k yields the following for $i \in \{1, \ldots, s\}$

$$g_k^i(\boldsymbol{x}) \leq A_k^{i-1} \sum_l a_{kl} |x_l|^i.$$

Therefore, we obtain that

$$\mathrm{E}\left(\sum_l a_{kl} |Y_{t-l}|\right)^i \leq A_k^{i-1} \sum_l a_{kl} \mathrm{E}|Y_{t-l}|^i \leq A_k^i \mathrm{E}|Y_0|^i. \tag{A-2}$$

Hence setting $\delta_{t;k} \equiv \sum_l a_{kl} |Y_{t-l}|$, we obtain that

$$\mathrm{E}\left(Y_t^s\right) \leq \sum_{k=1}^{K} p_k \mathrm{E}\lambda_{t;k}^s + \sum_{k=1}^{K} p_k \sum_{j=0}^{s-1} \left\{ \begin{matrix} s \\ j \end{matrix} \right\} \mathrm{E}[\lambda_{t;k}^j],$$

by using Lemma A-1. The last expression may be infinite but it is always well defined. Using repeatedly equations (A-1) and (A-2) and noting that the Y's are stationary, we obtain that

$$\mathrm{E}|Y_0|^s = \mathrm{E}\left(Y_t^s\right) \leq \sum_{k=1}^{K} p_k \mathrm{E}\delta_{t;k}^s + C_1 \leq B_s \mathrm{E}|Y_0|^s + C_2$$

for some constants $C_1 \leq C_2 < \infty$ since from recursion $\mathrm{E}|Y_t|^{s-1} < \infty$. Hence choose $B_s < 1$ to conclude that

$$\mathrm{E}|Y_0|^s \leq \frac{C_2}{1 - B_s} < \infty.$$

Step 3. If $s \notin \mathbb{N}^*$ then $S = [s] < s$ and we set $s = S + m$ for some $m \in (0, 1)$. Jensen's inequality entails that $B_S \leq B_s^{S/s} < 1$ thus the assumption also holds for S. Thus using step 2 for the integer value S and the sublinearity of $\lambda \mapsto \lambda^m$ we obtain that

$$f_k^s(\boldsymbol{x}) = f_k^S(\boldsymbol{x}) f_k^m(\boldsymbol{x}) \leq (g_k^S(\boldsymbol{x}) + r_k(\boldsymbol{x}))(g_k^m(\boldsymbol{x}) + c_k^m)) \leq g_k^s(\boldsymbol{x}) + R_k(\boldsymbol{x})$$

where the remainder term has degree $\leq S$, i.e. $R_k(\boldsymbol{x}) \leq C_k(\|\boldsymbol{x}\|^S + 1)$ for some finite constant $C_k < \infty$ and we thus obtain from step 2,

$$E\lambda_{k,0}^s \leq B_s E|Y_0|^s + C_k(E|Y_0|^S + 1)$$

and hence the desired conclusion follows as before. ■

A-3 Proof of Lemma 2.1

Because of the representation

$$Y_t = N_t(\lambda_t), \quad \lambda_t = h(\lambda_{t-1}, \ldots, \lambda_{t-L_1}, Y_{t-1}, \ldots, Y_{t-L_2})$$

we have that λ_t is measurable with respect to σ-field $\mathcal{F}_t$. By repeated substitution we obtain that there exists a positive function $f(\cdot)$ such that $\lambda_t = f(Y_{t-1}, Y_{t-2}, \ldots)$. Put $\delta = |\lambda_t - \tilde{\lambda}_t|$, and $\epsilon_t = |Y_t - \tilde{Y}_t|$. Then, by the contraction condition on $h(.)$, we obtain that

$$\delta_t \leq \sum_{i=1}^{L_1} a_i \delta_{t-i} + \sum_{j=1}^{L_2} \epsilon_{t-j}.$$

Define the polynomials $a(z) = 1 - \sum_{i=1}^{L_1} a_i z^i$ and $b(z) = \sum_{i=1}^{L_2} b_i z^i$. With this notation and because of positivity, the last display is written as

$$a(\mathcal{B})\delta_t \leq b(\mathcal{B})\epsilon_t$$

where $\mathcal{B}$ is the backward shift operator. Define $\nu_t = b(\mathcal{B})\epsilon_t - a(\mathcal{B})\delta_t \geq 0$. Since $\sum_{i=1}^{L_1} a_i < 1$, all the roots of $a(z)$ lie outside the unit circle. In case that $\nu_t = 0$, then

$$\delta_t = \frac{b(\mathcal{B})}{a(\mathcal{B})}\epsilon_t = \sum_{i=1}^{\infty} c_i |Y_{t-i} - \tilde{Y}_{t-i}|,$$

where the sequence (c_i) is positive and decays exponentially fast. In fact, by the multinomial expansion, it satisfies the following set of recursions

$$c_i = \sum_{j=1}^{L_1} b_j A_{i-j}, \;\; A_m = \sum_{l_1 + 2l_2 + \cdots + kl_k = m} \binom{l_1 + \cdots + l_k}{l_1, \ldots, l_k} a_1^{l_1} \cdots a_k^{l_k}.$$

In the special that $b(z) = 1$, the above derivation proves that the coefficients of the power series expansion of $1/a(z)$ are also positive. Indeed, when $b(z) = z$ the power series expansion of $z/a(z)$ has also positive coefficients and thus the claim follows. Therefore

$$\delta_t = \frac{b(\mathcal{B})}{a(\mathcal{B})}\epsilon_t - \frac{\nu_t}{a(\mathcal{B})} \leq \frac{b(\mathcal{B})}{a(\mathcal{B})}\epsilon_t.$$

The claim follows. ■

A-4 Proof of Lemma 3.1

The conclusion of the lemma will be true if we show that the functions

$$(y, y_1, \ldots, y_L) \mapsto \frac{\exp(-f(y_1, \ldots, y_L; \boldsymbol{\psi}_k)) f^y(y_1, \ldots, y_L; \boldsymbol{\psi}_k)}{y!},\ k \in \{1, \ldots, K\}$$

are linearly independent provided that $\boldsymbol{\psi}_k \neq \boldsymbol{\psi}_l$ for $k \neq l$. Indeed, let us assume that for all $y, y_1, \ldots, y_L$, a linear combination of these functions is equal to zero. Then

$$\sum_{k=1}^{K} p_k \frac{\exp(-f(y_1, \ldots, y_L; \boldsymbol{\psi}_k)) f^y(y_1, \ldots, y_L; \boldsymbol{\psi}_k)}{y!} = 0,$$

implies that

$$\sum_{k=1}^{K} p_k \exp(-f(y_1, \ldots, y_L; \boldsymbol{\psi}_k)) f^y(y_1, \ldots, y_L; \boldsymbol{\psi}_k) = 0.$$

But assumption **H-2** shows that for $k \neq l$, $f(.; \boldsymbol{\psi}_k) \neq f(.; \boldsymbol{\psi}_l)$. Hence, there exist a $(y_1, \ldots, y_L)$ and an index l exist such that, for $k \neq l$, $f(y_1, \ldots, y_L; \boldsymbol{\psi}_l) > f(y_1, \ldots, y_L; \boldsymbol{\psi}_k)$. By letting y going to infinity we deduce that $p_l = 0$. Hence

$$\sum_{k=1, k\neq l}^{K} p_k \exp(-f(y_1, \ldots, y_L; \boldsymbol{\psi}_k)) f^y(y_1, \ldots, y_L; \boldsymbol{\psi}_k) = 0,$$

and by recursion all p_k will be equal to zero. ■

A-5 Score and information matrix calculations for Theorem 3.1

Computation of first and second order derivatives: For ease of notation, define $\boldsymbol{y}_{(t-1)} = (y_{t-1}, \ldots, y_{t-L})$. By definition, we obtain that, for $k \in \{1, \ldots, K-1\}$,

$$\frac{\partial \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial p_k} = \frac{g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k) - g_K(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_K)}{g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}. \tag{A-3}$$

Now, for $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, d_k\}$ we get that

$$\frac{\partial \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki}} = \frac{p_k \Big(\partial g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k) / \partial \boldsymbol{\psi}_{ki} \Big)}{g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})},$$

with

$$\frac{\partial g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} = g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k) \frac{\partial f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} \left(\frac{y_t}{f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)} - 1 \right).$$

Therefore

$$\frac{\partial \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki}} = \frac{p_k g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)}{g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})} \frac{\partial f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} \left(\frac{y_t}{f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)} - 1 \right). \tag{A-4}$$

For the second order derivatives we get with $(k,l) \in \{1, \ldots, K-1\}^2$:

$$\frac{\partial^2 \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial p_k \partial p_l} = -\frac{\Big(g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}) - g_K(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})\Big)\Big(g_l(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}) - g_K(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})\Big)}{g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})^2}.$$

Similarly, for $k \in \{1, \ldots, K-1\}$ and $i \in \{1, \ldots, d_k\}$:

$$\frac{\partial^2 \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki} \partial p_k} = \frac{g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k) \frac{\partial f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} \left(\frac{y_t}{f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)} - 1 \right) g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})^2}$$

$$- \frac{p_k g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k) \frac{\partial f(\boldsymbol{y}_{(t-1))}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} \left(\frac{y_t}{f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)} - 1 \right) \Big(g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k) - g_K(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_K)\Big)}{g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})^2}.$$

For $(k,l) \in \{1, \ldots, K-1\}^2$, $k \neq l$ and $i \in \{1, \ldots, d_k\}$:

$$\frac{\partial^2 \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki} \partial p_l} = \frac{p_k g_k(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k) \frac{\partial f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} \left(\frac{y_t}{f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)} - 1 \right) \Big(g_l(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_l) - g_K(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_K)\Big)}{g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})^2}.$$

For $k \in \{1, \ldots, K-1\}$ and $(i,j) \in \{1, \ldots, d_k\}^2$:

$$\begin{aligned} \frac{\partial^2 \log g(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki} \partial \boldsymbol{\psi}_{kj}} &= p_k \frac{\partial^2 g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki} \partial \boldsymbol{\psi}_{kj}} \frac{1}{g(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi})} \\ &- p_k^2 \frac{\Big(\partial g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)/\partial \boldsymbol{\psi}_{ki}\Big)\Big(\partial g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)/\partial \boldsymbol{\psi}_{kj}\Big)}{g(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi})^2}, \end{aligned}$$

with

$$\begin{aligned} \frac{\partial^2 g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki} \partial \boldsymbol{\psi}_{kj}} &= \frac{\partial g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{kj}} \frac{\partial f(\boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} \Big(\frac{y_t}{f(\boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)} - 1 \Big) \\ &+ g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k) \left\{ \frac{\partial^2 f(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki} \partial \boldsymbol{\psi}_{kj}} \Big(\frac{y_t}{f(\boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)} - 1 \Big) - \frac{\partial f(\boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}} \frac{y_t \Big(\partial f(\boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)/\partial \boldsymbol{\psi}_{kj}\Big)}{f(\boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)^2} \right\} \end{aligned}$$

For $(k,l) \in \{1,\dots,K-1\}^2$, $k \neq l$, and $(i,j) \in \{1,\dots,d_k\} \times \{1,\dots,d_l\}$:

$$\frac{\partial^2 \log g(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki} \partial \boldsymbol{\psi}_{lj}} = -p_k p_l \frac{\left(\partial g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)/\partial \boldsymbol{\psi}_{ki}\right)\left(\partial g_k(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi}_k)/\partial \boldsymbol{\psi}_{kj}\right)}{g(y_t, \boldsymbol{y}_{t-1}; \boldsymbol{\psi})^2}$$

The special case of model (4): If $f(y_{t-1}, \dots, y_{t-L}; \boldsymbol{\psi}_k)$ is assumed to be a linear function like in equation (4), then the previous equations can be simplified considerably. More precisely

$$\begin{aligned}
\frac{\partial g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,0}} &= g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)\left(\frac{y_t}{\psi_{k,0} + \psi_{k,1} y_{t-1}} - 1\right) \\
\frac{\partial g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,1}} &= g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k) y_{t-1}\left(\frac{y_t}{\psi_{k,0} + \psi_{k,1} y_{t-1}} - 1\right).
\end{aligned}$$

For $k \in \{1,\dots,K-1\}$ and $i \in \{1,\dots,d_k\}$:

$$\begin{aligned}
\frac{\partial^2 \log g(y_t, y_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{k,0} \partial p_k} &= \frac{g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{g(y_t, y_{t-1}; \boldsymbol{\psi})}\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right) \\
&\quad - \frac{p_k g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right)\left(g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k) - g_K(y_t, y_{t-1}; \boldsymbol{\psi}_K)\right)}{g(y_t, y_{t-1}; \boldsymbol{\psi})^2} \\
\frac{\partial^2 \log g(y_t, y_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{k,1} \partial p_k} &= \frac{g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{g(y_t, y_{t-1}; \boldsymbol{\psi})}\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right) \\
&\quad - \frac{p_k g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k) y_{t-1}\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right)\left(g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k) - g_K(y_t, y_{t-1}; \boldsymbol{\psi}_K)\right)}{g(y_t, y_{t-1}; \boldsymbol{\psi})^2}
\end{aligned}$$

For $(k,l) \in \{1,\dots,K-1\}^2$, $k \neq l$ and $i \in \{1,\dots,d_k\}$:

$$\begin{aligned}
\frac{\partial^2 \log g(y_t, y_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{k,0} \partial p_l} &= \frac{p_k g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right)\left(g_l(y_t, y_{t-1}; \boldsymbol{\psi}_l) - g_K(y_t, y_{t-1}; \boldsymbol{\psi}_K)\right)}{g(y_t, y_{t-1}; \boldsymbol{\psi})^2} \\
\frac{\partial^2 \log g(y_t, y_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{k,1} \partial p_l} &= \frac{p_k g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k) y_{t-1}\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right)\left(g_l(y_t, y_{t-1}; \boldsymbol{\psi}_l) - g_K(y_t, y_{t-1}; \boldsymbol{\psi}_K)\right)}{g(y_t, y_{t-1}; \boldsymbol{\psi})^2}
\end{aligned}$$

For $k \in \{1,\dots,K-1\}$ and $(i,j) \in \{1,\dots,d_k\}^2$:

$$\begin{aligned}
\frac{\partial^2 g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,0} \partial \boldsymbol{\psi}_{k,0}} &= \frac{\partial g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,0}}\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right) - g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)\left(\frac{y_t}{\left(\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}\right)^2}\right) \\
\frac{\partial^2 g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,0} \partial \boldsymbol{\psi}_{k,1}} &= \frac{\partial g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,0}}\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right) - g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)\left(\frac{y_t y_{t-1}}{\left(\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}\right)^2}\right) \\
\frac{\partial^2 g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,1} \partial \boldsymbol{\psi}_{k,1}} &= \frac{\partial g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{k,1}} y_{t-1}\left(\frac{y_t}{\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}} - 1\right) - g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)\left(\frac{y_t y_{t-1}^2}{\left(\boldsymbol{\psi}_{k,0} + \boldsymbol{\psi}_{k,1} y_{t-1}\right)^2}\right)
\end{aligned}$$

Some remarks on Theorem 3.1 By construction, for any $k \in \{1, \ldots, K\}$, $|g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)| \leq 1$. In addition, $|g_k(y_t, y_{t-1}; \boldsymbol{\psi}_k)/g(y_t, y_{t-1}; \boldsymbol{\psi})| \leq 1$ and $|g(y_t, y_{t-1}; \boldsymbol{\psi})| \leq 1$. Therefore square integrability of the score function follows easily. Recall (A-3). Then

$$\Big|\frac{\partial \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial p_k}\Big| \leq 2.$$

Similarly, by (A-4) (with some abuse of notation)

$$\begin{aligned}\Big|\frac{\partial \log g(y_t, \boldsymbol{y}_{(t-1)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki}}\Big| &\leq \Big|\frac{\partial f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)}{\partial \boldsymbol{\psi}_{ki}}\Big|\Big|\left(\frac{y_t}{f(\boldsymbol{y}_{(t-1)}; \boldsymbol{\psi}_k)} - 1\right)\Big| \\ &\leq \Big(\sum_{l=1}^{L} b_{kli}|Y_l| + \Big|\frac{\partial f(\mathbf{0}, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}_{ki}}\Big|\Big)\frac{|Y_t|}{C_1},\end{aligned}$$

by using assumptions **H-2** and **H-4**. Square integrability of these score components follow by Cauchy-Schwartz and Theorem 2.1. Indeed, the score process is a square integrable martingale which satisfies Lindeberg's condition provided that $\mathrm{E}[Y_t^4] < \infty$. The martingale central limit theorem concludes the proof. ■

A-6 Verification of assumptions T1-T2 for linear MINARCH

We will illustrate, in the case of the simple linear model (4) assumptions **T1-T2** which are required in the proof of Theorem 3.2 and Proposition 3.1. Recall (4) where $(\psi_{0,k}, \psi_{1,k})_{1\leq k\leq K}$ are positive numbers, and for all $k \in \{1, \cdots, K\}$, $\psi_{0,k}$ greater or equal than a number $C_1 > 0$ (recall assumption **H-2**). We assume also that the true model satisfies

$$\mathrm{P}[Y_t = y\,|Y_{t-1}, \ldots Y_{t-L}] = \sum_{k=1}^{K} p_k^0 \frac{\exp(-(\psi_{k,0}^0 + \psi_{k,1}^0 Y_{t-1}))(\psi_{k,0}^0 + \psi_{k,1}^0 Y_{t-1})^y}{y!},$$

such that for all $k \in \{1, \cdots, K^0\}$, $0 \leq \psi_{k,1}^0 < 1$. Under these assumptions, we will show that there exists a $\delta > 0$ such that:

$$\sum_{y_{t-1}=0}^{\infty} e^{\delta y_{t-1}} \mu(y_{t-1}) < \infty, \tag{A-5}$$

where μ denotes the stationary measure.

Lemma A-2 Let $\psi_{1,1}^0, \cdots, \psi_{1,K^0}^0$ be the true regression parameters. If, for all $k \in \{1, \cdots, K^0\}$, $0 \leq \psi_{1,k}^0 < 1$, then there exists a $\delta_0 > 0$ such that for all $\delta < \delta_0$, (A-5) is true.

Proof: Assume that $Y_0 = 0$ and let $M > 0$. Then we have by the dominated convergence theorem:

$$\sum_{y_{t-1}=0}^{\infty} M \wedge e^{\delta y_{t-1}} \mu(y_{t-1}) = \lim_{t\to\infty} E\left(M \wedge e^{\delta Y_t}\right).$$

Let $\rho_0 = \max_{1\leq k\leq K^0}(\psi_{0,k}^0)$ and $\rho_1 = \max_{1\leq k\leq K^0}(\psi_{1,k}^0)$. Observe that if $Z_t \sim \mathrm{Poisson}(\rho_1 Z_{t-1} + \rho_0)$ and $Z_0 = Y_0$, then, for all $t \in \mathbb{N}$

$$\mathrm{E}\left(e^{\delta Y_t}\right) \leq \mathrm{E}\left(e^{\delta Z_t}\right).$$

But $\mathrm{E}\left(\exp(\delta Z_1)\right) = \exp(\rho_0(e^{\delta}-1))$ and $\mathrm{E}\left(\exp(\delta Z_2)\right) = \exp(\rho_0(e^{\delta}-1))\exp(\rho_0\left(e^{\rho_1(e^{\delta}-1)}-1\right))$. Moreover, for $\rho_1 < C < 1$ and if $\delta < \ln\left(C/\rho_1\right)$ we obtain

$$\mathrm{E}\left(\exp(\delta Z_2)\right) \leq \exp(\rho_0(e^{\delta}-1)(1+C)).$$

By recursion and taking the limit as $t \to \infty$

$$\lim_{t\to\infty} \mathrm{E}\left(\exp(\delta Z_t)\right) \leq \exp(\rho_0(e^{\delta}-1)(\frac{1}{1-C})).$$

Therefore the claim follows by noting that

$$\lim_{M\to\infty} \sum_{y_{t-1}=0}^{\infty} M \wedge e^{\delta y_{t-1}}\mu(y_{t-1}) = \sum_{y_{t-1}=0}^{\infty} e^{\delta y_{t-1}}\mu(y_{t-1}) \leq e^{\rho_0(e^{\delta}-1)(\frac{1}{1-C})}$$

■

Verification of Assumption T-1 for model (4): Assumption **T1** is a consequence of assumptions **H1**, **H-2** and **H-3**. Assumption **H1** has to be assumed. The assumption **H-2** follows from the linearity and that and for all $y \in \mathbb{N}$, $P(Y_t = y) > 0$ if Y_t is stationary. Finally, since for all $k \in \{1,\ldots,K\}$, $p_k \geq C > 0$ and $\exp\left[-(\psi_{0,k}+\psi_{1,k}Y_{t-1})\right](\psi_{0,k}+\psi_{1,k}Y_{t-1})^{y_t}/y_t! \leq 1$, we have

$$\begin{aligned}
\log\left(g(y_t, y_{t-1}; \boldsymbol{\psi})\right) &= \log\left(\sum_{k=1}^{K} p_k \frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!}\right) \\
&\leq \log(C) + \sum_{k=1}^{K} \log \frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!} \\
&= \log(C) + \sum_{k=1}^{K} -(\psi_{0,k}+\psi_{1,k}y_{t-1}) + y_t \log(\psi_{0,k}+\psi_{1,k}y_{t-1}) + \log(y_t!) \\
&\leq \log(C) + \sum_{k=1}^{K} \left[-(\psi_{0,k}+\psi_{1,k}y_{t-1}) + y_t \log(\psi_{0,k}+\psi_{1,k}y_{t-1})\right] + y_t \log(y_t).
\end{aligned}$$

Because of the compactness of the parameter space, we obtain

$$\mathrm{E}\left(\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} g(y_t, y_{t-1}; \boldsymbol{\psi})\right) < \infty$$

provided $\mathrm{E}(Y_t^s) < \infty$ for some $s > 1$.

Verification of Assumption T-2 for model (4): After some calculations, we obtain that

$$\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)}^2 = \sum_{y_t=0}^{\infty} \sum_{y_{t-1}=0}^{\infty} \frac{\left(\sum_{k=1}^{K} p_k \frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!}\right)^2}{\sum_{k=1}^{K^0} p_k^0 \frac{\exp -(\psi_{0,k}^0+\psi 0_{1,k}y_{t-1})(\psi_{0,k}^0+\psi_{1,k}^0 y_{t-1})^{y_t}}{y_t!}} \mu\left(y_{t-1}\right) - 1.$$

By employing the inequality

$$\left(\sum_{k=1}^{K} p_k \frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!}\right)^2 \leq \sum_{k=1}^{K} p_k \left(\frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!}\right)^2,$$

the integral will be finite if, for all $k \in \{1,...,K\}$:

$$\sum_{y_t=0}^{\infty}\sum_{y_{t-1}=0}^{\infty} \frac{\left(\frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!}\right)^2}{\sum_{k=1}^{K^0} p_k^0 \frac{\exp -(\psi_{0,k}^0+\psi_{1,k}^0 y_{t-1})(\psi_{0,k}^0+\psi_{1,k}^0 y_{t-1})^{y_t}}{y_t!}} \mu\left(y_{t-1}\right) < \infty$$

On the other hand, since

$$\sum_{k=1}^{K^0} p_k^0 \frac{\exp -(\psi_{0,k}^0+\psi 0_{1,k}y_{t-1})(\psi_{0,k}^0+\psi_{1,k}^0 y_{t-1})^{y_t}}{y_t!} \geq p_k^0 \frac{\exp -(\psi_{0,k}^0+\psi 0_{1,k}y_{t-1})(\psi_{0,k}^0+\psi_{1,k}^0 y_{t-1})^{y_t}}{y_t!}$$

for every $k \in \left\{1,...,K^0\right\}$, the generalized score function is well defined provided that for every $k \in \{1,...,K\}$, there exists $l \in \left\{1,...,K^0\right\}$ such that

$$\sum_{y_t=0}^{\infty}\sum_{y_{t-1}=0}^{\infty} \frac{\left(\frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!}\right)^2}{\frac{\exp -(\psi_{0,l}^0+\psi_{1,l}^0 y_{t-1})(\psi_{0,l}^0+\psi_{1,l}^0 y_{t-1})^{y_t}}{y_t!}} \mu\left(y_{t-1}\right) < \infty.$$

But the Poisson pmf is always less than 1 and therefore this will be true if

$$\sum_{y_t=0}^{\infty}\sum_{y_{t-1}=0}^{\infty} \frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!} \frac{\frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!}}{\frac{\exp -(\psi_{0,l}^0+\psi_{1,l}^0 y_{t-1})(\psi_{0,l}^0+\psi_{1,l}^0 y_{t-1})^{y_t}}{y_t!}} \mu\left(y_{t-1}\right) < \infty,$$

or if

$$\sum_{y_t=0}^{\infty}\sum_{y_{t-1}=0}^{\infty} \left(\exp(\psi_{0,l}^0-\psi_{0,k})\exp((\psi_{1,l}^0-\psi_{1,k})y_{t-1}\right) \frac{\exp -(\psi_{0,k}+\psi_{1,k}y_{t-1})(\psi_{0,k}+\psi_{1,k}y_{t-1})^{y_t}}{y_t!} \exp\left(y_t \log\left(\frac{\psi_{0,k}+\psi_{1,k}y_{t-1}}{\psi_{0,l}^0+\psi_{1,l}^0 y_{t-1}}\right)\right) \mu\left(y_{t-1}\right) < \infty.$$

By summing with respect to y_t, and recognizing the moment-generating function of a Poisson random variable the previous inequality becomes:

$$\sum_{y_{t-1}=0}^{\infty} \exp(\psi_{0,l}^0-\psi_{0,k})\exp((\psi_{1,l}^0-\psi_{1,k})y_{t-1})\exp((\psi_{0,k}+\psi_{1,k}y_{t-1})\left(\frac{\psi_{0,k}+\psi_{1,k}y_{t-1}}{\psi_{0,l}^0+\psi_{1,l}^0 y_{t-1}}-1\right)\mu\left(y_{t-1}\right) < \infty.$$

Since the mixtures weigths are bounded by below, if $\|g-g^0\|$ is sufficiently small, then $\min_{(k,l)\in\{1,\cdots,K\}\times\{1,\cdots,K^0\}}((\psi_{0,l}^0-\psi_{0,k})^2+(\psi_{1,l}^0-\psi_{1,k})^2)$ is sufficiently small , and the last inequality will be true as soon as, a $\delta > 0$ exists such that:

$$\sum_{y_{t-1}=0}^{\infty} \exp(\delta y_{t-1})\mu(y_{t-1}) < \infty$$

But this follows from Lemma A-2.

References

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

Albert, P. S. (1991). A two–state Markov mixture model for a time series of epileptic seizure counts. *Biometrics 47*, 1371–1381.

Berentsen, G. D., J. Bulla, A. Maruotti, and B. Støve (2018). Modelling corporate defaults: A Markov-switching Poisson log-linear autoregressive model. *ArXiv e-prints*. https://arxiv.org/abs/1804.09252.

Berkes, I., L. Horváth, and P. Kokoszka (2003). GARCH processes: structure and estimation. *Bernoulli 9*, 201–227.

Carvalho, A. X. and M. A. Tanner (2005). Modeling nonlinear time series with local mixtures of generalized linear models. *Canad. J. Statist. 33*, 97–113.

Carvalho, A. X. and M. A. Tanner (2007). Modelling nonlinear count time series with local mixtures of Poisson autoregressions. *Comput. Statist. Data Anal. 51*, 5266–5294.

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

Davis, R. A., S. H. Holan, R. Lund, and N. Ravishanker (Eds.) (2016). *Handbook of Discrete-Valued Time Series*. Handbooks of Modern Statistical Methods. London: Chapman & Hall/CRC.

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

Dedecker, J. and C. Prieur (2004). Coupling for τ-dependent sequences and applications. *Journal of Theoretical Probability 17*, 861–885.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38. with discussion.

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

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

Doukhan, P., K. Fokianos, and X. Li (2012). On weak dependence conditions: The case of discrete valued processes. *Statistics & Probability Letters 82*, 1941 – 1948.

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

Doukhan, P. and W. Kengne (2015). Inference and testing for structural change in general Poisson autoregressive models. *Electronic Journal of Statistics 9*, 1267–1314.

Doukhan, P. and M. H. Neumann (2017). Absolute regularity of semi-contractive GARCH-type processes. *ArXiv e-prints*. https://arxiv.org/abs/1711.04282.

Doukhan, P. and O. Wintenberger (2008). Weakly dependent chains with infinite memory. *Stochastic Processes and Their Applications 118*, 1997–2013.

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

Fokianos, K. and R. Fried (2010). Interventions in INGARCH processess. *Journal of Time Series Analysis 31*, 210–225.

Fokianos, K. and M. Pitsillou (2018). Testing independence for multivariate time series via the auto-distance correlation matrix. *Biometrika* **105**, 337–352.

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

Fokianos, K. and D. Tjøstheim (2012). Nonlinear Poisson autoregression. *Annals of the Institute of Statistical Mathematics 64*, 1205–1225.

Francq, C. and J.-M. Zaköian (2001). Stationarity of multivariate Markov-switching ARMA models. *J. Econometrics 102*, 339–364.

Francq, C. and J.-M. Zakoïan (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli 10*, 605–637.

Jacobs, R., M. Jordan, S. Nowlan, , and G. Hinton (1991). Adaptive mixtures of local experts. *Neural Computation 3*, 79–87.

Johnson, N. L., S. Kotz, and A. W. Kemp (1992). *Univariate Discrete Distributions* (second ed.). New York: Wiley.

Kedem, B. and K. Fokianos (2002). *Regression Models for Time Series Analysis*. Hoboken, NJ: Wiley.

Le, N. D., R. D. Martin, and A. E. Raftery (1996). Modelling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association 91*, 1504–1515.

Liboschik, T., K. Fokianos, and R. Fried (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software 82*.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

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

Olteanu, M. and J. Rynkiewicz (2012). Asymptotic properties of autoregressive regime-switching models. *ESAIM Probab. Stat. 16*, 25–47.

Rydberg, T. H. and N. Shephard (2000). 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 (Eds.), *Nonlinear and Nonstationary Signal Processing*, pp. 217–246. Cambridge: Isaac Newton Institute and Cambridge University Press.

Saikkonen, P. (2007). Stability of mixtures of vector autoregressions with autoregressive conditional heteroskedasticity. *Statist. Sinica 17*, 221–239.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

Taniguchi, M. and Y. Kakizawa (2000). *Asymptotic theory of statistical inference for time series*. New York: Springer.

Tjøstheim, D. (2012). Some recent theory for autoregressive count time series. *TEST 21*, 413–438.

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

Truquet, L. (2017). A perturbation analysis of some Markov chains models with time-varying parameters. *ArXiv e-prints*. https://arxiv.org/abs/1706.03214.

Wong, C. S. and W. K. Li (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society B62*, 95–115.

Wong, C. S. and W. K. Li (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association 96*, 982–995.

Wood, S., O. Rosen, and R. Kohn (2011). Bayesian mixtures of autoregressive models. *J. Comput. Graph. Statist. 20*, 174–195.

Zhu, F., Q. Li, and D. Wang (2010). A mixture integer-valued ARCH model. *J. Statist. Plann. Inference 140*, 2025–2036.